2012

What's in a letter?

Aaron J Schein

What's in a Letter?

A Thesis Presented

by

AARON JOSEPH STERIADE SCHEIN

Submitted to the Graduate School of the

University of Massachusetts in partial fulfillment

of the requirements for the degree of

MASTER OF LINGUISTICS

May 2012

LINGUISTICS

ABSTRACT

What's in a Letter?

May 2012

AARON JOSEPH STERIADE SCHEIN

B.A., University of Massachusetts Amherst

M.A., University of Massachusetts Amherst

Directed by: Professor Brian Dillon

Sentiment analysis is a burgeoning field in natural language processing used to extract and categorize opinion in evaluative documents. We look at recommendation letters, which pose unique challenges to standard sentiment analysis systems. Our dataset is eighteen letters from applications to UMass Worcester Memorial Medical Center's residency program in Obstetrics and Gynecology. Given a small dataset, we develop a method intended for use by domain experts to systematically explore their intuitions about the topical make-up of documents on which they make critical decisions. By leveraging WordNet and the WordNet Propagation algorithm, the method allows a user to develop topic seed sets from real data and propagate them into robust lexicons for use on new data. We show how one pass through the method yields useful feedback to our beliefs about the make-up of recommendation letters. At the end, future directions are outlined which assume a fuller dataset.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Section 1: Introduction

Many on the front lines of text-mining and natural language processing today are working to develop systems for extracting opinion and subjectivity in text. The expression of opinion in natural language is highly nuanced and domain-specific which makes the development of such systems very difficult. Systems developed for use in one domain do not usually function accurately in others.

Subjectivity is found universally – but its detection and categorization is only useful in some areas. In general, opinion extraction systems are developed to analyze collections of documents which are inherently evaluative, product reviews being the most the common.

A year and half ago, a doctor and professor who serves as director of the residency program in obstetrics and gynecology at UMass Worcester's Memorial Medical Center, contacted the Linguistics department with a simple question: could we empirically measure the strength of recommendation letters? As he explained, the residency program admissions board spends weeks every year pouring over hundreds of reference letters debating the significance and meaning of certain words and phrases. Was there a scientific way to resolve disputes over whether a certain phrase meant, for example, that an applicant was truly exceptional?

We began this research in January 2011 first by discussing how to compile a dataset and what the residency program wanted in an opinion-extraction system. This paper outlines a method we tried for testing intuitions about the topical make-up of reference letters given a very limited dataset, which is what we ultimately had. We also spent time outlining which steps take once a fuller dataset became available – these future directions are mentioned at the end.

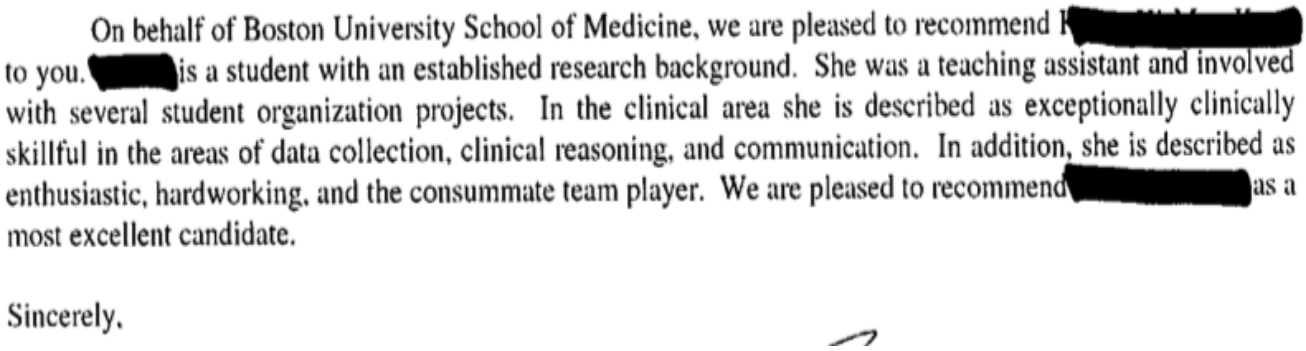
Section 2: Data

2.1. Letters of recommendation

Applications to residency programs contain letters of recommendation written by professors, doctors, and administrators from the applicant's medical school. The form and number of an applicant's letters vary depending on their school of origin. For this study, we limited our data to letters for

applicants applying from Boston University School of Medicine. The standard application from BU contains three reference letters written by professors and doctors who have had extended contact with the applicant. It also contains the Dean's Letter, written by the Dean's Committee, the form of which is standardized. The Dean's Letter initially seemed like the right point of entry because it varies in specific and predictable places. The following Figures 1 and 2 show side-by-side the "Summary" section of two candidate's Dean's letters.

Figure 1. "Summary" section of Candidate 6's Dean's Letter

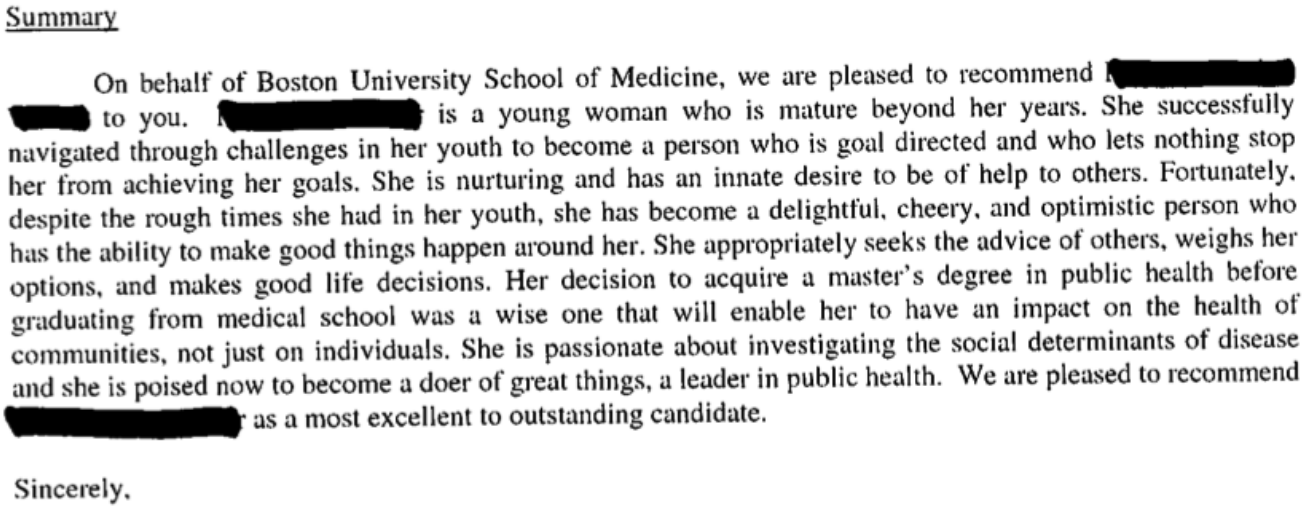


Summary

On behalf of Boston University School of Medicine, we are pleased to recommend [redacted] to you. [redacted] is a student with an established research background. She was a teaching assistant and involved with several student organization projects. In the clinical area she is described as exceptionally clinically skillful in the areas of data collection, clinical reasoning, and communication. In addition, she is described as enthusiastic, hardworking, and the consummate team player. We are pleased to recommend [redacted] as a most excellent candidate.

Sincerely,

Figure 2. "Summary" section of Candidate 3's Dean's Letter



Summary

On behalf of Boston University School of Medicine, we are pleased to recommend [redacted] to you. [redacted] is a young woman who is mature beyond her years. She successfully navigated through challenges in her youth to become a person who is goal directed and who lets nothing stop her from achieving her goals. She is nurturing and has an innate desire to be of help to others. Fortunately, despite the rough times she had in her youth, she has become a delightful, cheery, and optimistic person who has the ability to make good things happen around her. She appropriately seeks the advice of others, weighs her options, and makes good life decisions. Her decision to acquire a master's degree in public health before graduating from medical school was a wise one that will enable her to have an impact on the health of communities, not just on individuals. She is passionate about investigating the social determinants of disease and she is poised now to become a doer of great things, a leader in public health. We are pleased to recommend [redacted] as a most excellent to outstanding candidate.

Sincerely,

These short "Summary" sections are the most free-form sections of the Dean's Letter and yet, their

form still conforms to certain consistent standards. Form consistency enables opinion extraction. When a set of evaluative documents are structured similarly, a system can find the subjective parts dependably. Accurately classifying subjectivity is a prerequisite for extracting opinion and is itself a non-trivial problem on unstructured text.[1]

In the end, we decided to study the letters reference and not the Dean's Letter. The reason for this can be found in the last sentence of the "Summary": "We are pleased to recommend _____ as a _____ candidate." This last sentence is one example by which the Dean's Letter is a formula. The adjectives modifying "candidate" – "most excellent to outstanding" for Candidate 6 and "most excellent" for Candidate 3 – refer to a literal ranking which the letter explicitly announces in an earlier section, seen below in Figure 3.

Figure 3. Boston University School of Medicine's ranking policy

Boston University School of Medicine does not rank its students, however, we give an overall statement of recommendation for each student in the final paragraph. We believe that the overwhelming majority of our students have prepared themselves to be truly excellent physicians, and will perform at a very high level in their residency programs. The statement of recommendation reflects this belief, and it also calls attention to those students who have distinguished themselves at an outstanding level. The statement of recommendation includes Very Good, Excellent, Most Excellent, Most Excellent to Outstanding, Outstanding, and Most Outstanding (top 12 % of the Class of 2011). The number of students described as Most Outstanding is similar to previous years.

These terms, like "Most Excellent", are code referring to different percentiles of the student body, a subtle way of providing a ranking to residency programs without an overt ranking policy. The Dean's Committee describes their perspective as the "10,000 foot view". The main information encoded in the Dean's Letter is a relative ranking of the applicant based on their grades, scores, and their reference letters. Since their exact grades and scores are available to the residency admissions board, we decided that analyzing only the reference letters was the most direct way to add value to the admissions process.

Section 2.2. Limitations to compiling a dataset

[1] J. M. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning subjective language," Computational Linguistics, vol. 30, pp. 277–308, September 2004

For issues of privacy, letters of recommendation are submitted to the residency program admissions board through an online database that makes available read-only PDFs for download. For this reason, obtaining digital copies of the letters was difficult because they first had to be redacted. Since, read-only PDFs do not allow editing, individual staff members had to print out hard-copies of the letters, and black-out all instances of the candidates' names. The redacted versions were then scanned and sent to us as picture format files. We used optical character recognition software and manual typing to convert these into text files. There were only two staff members at the residency program who had permission to view the non-redacted letters. Due to their busy schedules, we were only able to produce a dataset of eighteen letters – three letters for six candidates. Sample data is available in Appendix A.

Section 3: Method

Section 3.1. Sentiment analysis overview

Sentiment analysis or opinion mining refers to a collection of natural language processing techniques succinctly defined as, "the computational treatment of opinion, sentiment, and subjectivity in text."[2] This area of research has enjoyed a surge in popularity as the volume of opinion-rich resources (e.g. blogs, reviews) has increased:

> "What other people think has always been an important piece of information for most of us during the decision-making process . . . many of us asked our friends to recommend an auto mechanic or to explain who they were planning to vote for in local elections, requested reference letters regarding job applicants from colleagues, or consulted Consumer Reports to decide what dishwasher to buy."[3]

Opinion is as important to companies, interested in measuring the success of their products, as it is to

[2] "Opinion Mining and Sentiment Analysis", 2008, B. Pang, L. Lee, 1
[3] Ibid., 8

consumers. It is important for political campaigns and intelligence analysts. Similarly, opinion is important to admissions boards that seek insight beyond the quantitative metrics provided by grades and standardized testing.

A benchmarking dataset for sentiment analysis systems is the movie review corpus.[4] This corpus comprises several thousand reviews each coupled with a five star rating. The goal of a sentiment analysis system given the movie review corpus is to accurately classify the five star rating of a review based on natural language features in the corresponding text. The longer term hope for these systems is that they discover domain-specific natural language features that accurately predict these various metrics. In cases where quantitative metrics are not provided, a successful system is able to quantify and categorize opinion on a large scale and summarize it in a meaningful way.

Recommendation letters, like movie reviews, come paired with corresponding quantitative features. An application for residency includes an applicant's GPA in medical school and university, MCAT scores, and class ranking. A possible approach to the recommendation letters problem could be to predict an applicant's GPA, MCAT, or some other quantitative feature given the text of their letters. If developed, such a system could be used by admissions boards to flag mismatches: applicants who have low GPAs but letters written about them as if they had high ones.[5] However, the usefulness and validity of this approach seems limited. Admissions boards will always be provided with an applicant's exact quantitative scores. Besides flagging mismatches, the information provided by a system which predicts those scores is redundant. Moreover, the point of reference letters is to provide information about the applicant that cannot be provided by their scores. Developing systems which can automatically categorize and summarize this information would be much more useful.

Sentiment analysis systems are generally lexicon-based, meaning they depend, in part, on the development of sentiment lexicons: bins of terms corresponding to one of the sentiment classifications. Most systems recognize two opposing categories, positive and negative, and a third neutral one. Some

[4]http://www.cs.cornell.edu/people/pabo/movie-review-data/
[5]Or, alternatively, the opposite.

systems distinguish different types of positive or negative sentiment like, "happy" and "excited" or "sad" and "angry".[6] These lexicons can be manually hand-picked by domain experts or inducted through supervised or unsupervised learning techniques and through bootstrapping methods.[7] Once sentiment lexicons have been developed, a system classifies documents according to some function of the frequency of the terms from each lexicon in the document. The most basic system is a binary classifier which classifies a document as positive or negative if the terms from one lexicon appear more frequently in the document than the terms from the other.

Section 3.2. Challenges

Lexicon-based approaches on the simple unigram model have achieved over 80% accuracy on the movie review corpus.[8] However, recommendation letters unique challenges. Unlike movie reviews, recommendation letters do not feature the full range of human emotion. A reference letter is inherently positive – its purpose is not just for reviewing an applicant but recommending them. Sentiment in letters ranges from positive to exceptionally positive, the distinction being highly nuanced and more difficult to detect with the standard approach. Moreover, letter writing is strategic. A letter writer may be reserved about giving an applicant their best recommendation, but will take measures to obscure and qualify their true opinion. While the purpose of movie and product reviews is to telegraph an honest opinion to prospective consumers, the purpose of reference letters is to highlight the best aspects of an applicant.

One way in which the recommendation letters problem may be simpler than movie or product reviews should be noted. Movie or product review algorithms often suffer in accuracy from sarcasm (e.g. "that movie was just SO amazing"), hypallage (e.g. "dialogue was not exactly brilliant"), or

[6] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction," in Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005.

[7] "Opinion Mining and Sentiment Analysis", 2008, B. Pang, L. Lee, 44

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86, 2002

hyperbole (e.g. “literally the worst movie in the history of the world”).[9] Letters of reference, on the other hand, will likely keep to formal prose.

Section 3.3. Topic lexicons

We implemented a different approach to sentiment analysis in an attempt to bypass the challenges posed by reference letters. Instead of designing a system for the automatic detection of sentiment – which we assumed would always be positive – we tried using lexicon-based techniques to detect the mixture of topics in a reference letter. While a letter will always recommend a candidate, it may not always recommend the same characteristics of a candidate. We found this was an intuition shared by the residency program staff and a number of professionals and academics we conferred with.

The state-of-the-art in sentiment analysis today are systems that can identify entities in a document, for example, “iPhone”, extract features of the entity, like “touchscreen” or “weight”, and analyze the sentiment towards those features. Instead of producing vague summaries like “Positive” or “Negative”, these systems can map specific features of entities to sentiment, like “Positive about the touchscreen” or “Negative about the weight”. As Bing Liu puts it, these cutting-edge systems produce “actionable” information, unlike their simpler counterparts.[10]

We assume a recommendation letter will feature positive sentiment throughout. So the task is to identify which specific features or characteristics of the applicant are being recommended. Such a system could be used in a number of ways. It could be used to discover which topics or characteristics are correlated with the applicant's quantitative metrics. Perhaps, for example, high GPAs are highly correlated with letters that focus heavily on an applicant's intellectual ability. Moreover, not all residency programs look for the same characteristics in an applicant. Such a system could also be used by residency programs to find applicants who are recommended for certain desired characteristics, like their ability to communicate and work well with others.

Instead of developing sentiment lexicons, we developed topic lexicons – bins of words which

[9] C. Potts, “Sentiment Symposium Tutorial,” in Sentiment Analysis Symposium, San Francisco, 2011
[10] B. Liu. (2012, May). *Sentiment Analysis Tutorial*, Speech presented at Sentiment Analysis Symposium, New York, NY.

we believed corresponded with a specific characteristics a letter writer might recommend. Our lack of sufficient data prevented us from implementing any machine learning methods for lexicon induction. To develop the topic lexicons, we leveraged ours and others' intuitions about the topical make-up of reference letters, and defined a set of general topics which we believed would account for a large proportion of the tokens in the letters. Having defined these topics, we carefully read through each of the letters in our dataset. When we came across a token in the letters that we believed signified a certain topic, we added it to that topic's lexicon. Having compiled these initial lexicons, we used WordNet to propagate and enhance them.

Section 3.4. WordNet

Throughout government, business, and academia, a host of domains benefit from lexicon-based tools. Entity extraction and resolution, for example, is a ubiquitous and crucial function for projects which depend on high quantities of natural language data. Depending on a project's domain, the terms, for example, “New York City” and “New York” may or may not be resolved to the same entity. These choices are highly contextual and often require the development of entity gazetteers which map certain expected terms to entities. Lexicons which are developed manually by domain experts can be highly nuanced but are often insufficiently robust. A series of unsupervised bootstrapping methods exist for lexicon induction.[11]However, like most unsupervised methods, these require a large amount of data.

For projects dealing with nuanced natural language and but lacking the volume of data necessary for powerful lexicon induction methods, an alternative exists for enhancing manually created lexicons. WordNet is a database of lexical relations in English which distinguishes the various senses of words. The senses of words are grouped into “cognitive synonyms” known as synsets.[12] For example, WordNet distinguishes two senses of word “leader”. These senses, both nouns, are members of two separate synsets, seen below with their corresponding WordNet definitions:

[11] E. Rilo and J. Wiebe, “Learning extraction patterns for subjective expressions,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

[12] http://wordnet.princeton.edu/

1. 'leader.n.01': 'a person who rules or guides or inspires others'

2. 'drawing_card.n.02': 'a featured article of merchandise sold at a loss in order to draw customers'

A synset is linked to other synsets through a series of possible relations depending on its part-of-speech. Between synsets of nouns, for example, WordNet maps hypernyms, hyponyms, coordinate terms, holonyms, and meronyms.[13] For our project, we leveraged synset relations which encoded semi-synonymous and antonymous links.

Section 3.5. WordNet Propagation Algorithm

Algorithms exist to propagate lexicons by leveraging WordNet's lexical mappings. Foremost among these is the WordNet Propagation algorithm, developed for sentiment analysis by Chris Potts.[14] A graphical representation is found in Appendix B. The algorithm finds words with similar senses to those in a seed lexicon and appends them. It is designed to preserve the semantic properties of some hand-selected lexicons, known as seed sets, by "[traveling] strategically through WordNet"[15]. A user of the algorithm passes it two seed sets of opposing semantic properties and a desired number of iterations. A third seed set for neutral or objective terms is optional. Through each iteration, the algorithm finds the direct synonyms and antonyms of each synset in the seed sets. For each of the two opposing seed sets, the algorithm appends the synonyms of that set's synsets and the antonyms of the opposing seed set's synsets. Pruning overlap is an an option specified by the user. When this is in effect, a discovered synset is not appended if it already exists in the opposing seed set, prioritizing more direct lexical relations.

Semantic opposition of the seed sets is an important aspect of WordNet Propagation. Developed specifically for sentiment analysis, the semantic properties of the seed sets are generally some variation of "Positive", "Negative", and "Neutral". However, many other sets of polarized

[13] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
[14] http://sentiment.christopherpotts.net/lexicons.html#wnpropagate
[15] Ibid.

classes are possible. Chris Potts provides examples of seeds sets in his tutorial on WordNet – they are below in Tables 1 and 2.

Table 1. Example of standard opposing seed sets.[16]

Category	Seed set
"Positive"	Excellent, good, nice, positive, fortunate, correct, superior
"Negative"	Nasty, bad, poor, negative, unfortunate, wrong, inferior
"Objective"	Administrative, financial, geographic, constitute, analogy, ponder, material, public, department, measurement, visual

Table 2. Example of non-standard opposing seed sets.[17]

Category	Seed set
"Pleasure"	Amuse, calm, ecstasy, enjoy, joy
"Pain"	Agony, disconcerted, fearful, regret, remorse
"Strong"	Illustrious, rich, control, perseverance
"Weak"	Lowly, poor, sorry, sluggish, weak
"Male"	Boy, brother, gentleman, male, guy
"Female"	Girl, sister, bride, female, lady

In classifying recommendation letters, for the reasons discussed in <SECTION>, the standard application of WordNet Propagation is inappropriate. The difference between an average versus above-average applicant could manifest itself in the appearance of positive adjectives versus superlative adjectives (e.g. "good" versus "best"). Alternatively, an average applicant may be described as exhibiting a subset of some set of desired features while an above-average applicant may be described as exhibiting a larger subset. In either case, semantically opposed lexicons will fail to capture this reality: the words "good" and "best" are not opposites, nor "driven" and "intelligent".

Section 3.6. Adapting WordNet Propagation

We used a simple trick to leverage WordNet Propagation's power for our purposes. While we were interested in enhancing the topic lexicons, we were not interested in developing opposing lexicons.

[16] Ibid.
[17] http://sentiment.christopherpotts.net/lexicons.html#wnpropagate

For example, the “Professional” lexicon we hand-picked, contained the words “professional” and “motivated”. Their opposites exist – “unprofessional” and “unmotivated” – however, we do not expect to discover such overtly negative terms in letters of recommendation. Although the algorithm works through semantic opposition of the seed sets, it does set any requirement that the seed sets are of similar size. So, for each topic lexicon, we passed WordNet Propagation a seed set for the lexicon and an empty seed set as its opposing lexicon. In the first iteration, the algorithm appended the synonyms of the synsets from the topic lexicon and appended their antonyms to the opposing seed set. Since in the first iteration the opposing seed set was empty, the algorithm found no synonyms or antonyms and moved on. In the second and subsequent iterations, the algorithm worked as usual – finding synonyms and antonyms of both opposing seed sets and propagating both. We experimented with a range of iterations, the discussion of which is described in <SECTION>. After it completed, we used only the set we were interested in, the propagated topic lexicon, and discarded the opposing set.

Section 3.7. Seed set selection

We spent some time with the data, reading letters, developing our own intuitions about their general make-up and discussing them with the residency admissions staff. Eventually, we developed what we believed to be a set of topics which approximated an exhaustive partition of the possible topics and sub-topics in letters of reference for applicants to residency programs. The topics are briefly described here and the initial seed sets are available in Appendix C.

“Professional”: These terms correspond to a candidate's maturity, organizational habits, time-management skills, and professional demeanor. In general, these were terms which described the process of a candidate's work but not the work itself or its outcomes. Examples are “hard-working”, “efficient”, “mature”.

“Skill”: These terms correspond to a candidate's intellectual ability and skill set. Examples are “sharp”, “adept”, “knowledgeable”.

“Teamwork”: These terms correspond to a candidate's ability to articulate their ideas, communicate effectively, and work well in groups. The residency staff stressed on a number of occasions that a candidate's ability to perform well in residency was highly dependent on their ability to collaborate with others in fast-paced environments. We found that many letters focused on this characteristic. Examples are “articulate”, “collaborative”, “leadership”.

“Personal”: These terms correspond to a candidate's general personality and likability. This topic was vague, but appropriately so. In a way, this topic serves as a proxy for filler. We found that letters which devoted less space to stating vaguely positive personal characteristics were stronger whereas letters that devoted more space to reporting about a candidate's general likability were weaker. Examples are “bright”, “energetic”, “lovely”.

“Compassion”: These terms correspond to a candidate's conscience, ethics, sense of purpose, and their commitment to social justice. There was overlap between this and the “Personal” topic but we found it was necessary to distinguish them. Certain letters focused on a candidate's volunteer work or involvement and leadership in social justice movements. These often also focused on a candidate's general sense of mission and motivation to better the world. We found that letters either focused on this topic heavily or didn't feature it at all. Examples are “compassionate”, “ethical”, “socioeconomic”.

“Superlative”: This topic was for strong qualifying adverbs and superlative adjectives. Examples are “unquestionably”, “best”, “most”.

Section 3.8. Scoring

The goal of propagating our topics was to develop robust lexicons which signaled certain properties corresponding to frequently mentioned candidate features. To capture the topics' prevalence in a letter we measured two scores, the proportion and frequency of each lexicon's terms in the letter. These two scores allowed us to compare topic prevalence within a candidate and between candidates respectively. Normalizing the scores by the length of a letter (proportion) allows for statements about a

candidate's strengths and weaknesses (within candidate). For example, if 10% of a letter's tokens are positive hits for the “Professional” topic and 25% are hits for the “Skill” topic, there is evidence that the candidate is being recommended more for their skill and ability than for their professionalism. Unnormalized scores (frequency) allow for comparisons between candidates. If Candidate A's letters have 75 hits for “Skill” and 62 hits for “Professional” while Candidate B's letters have 150 hits for “Skill” and 140 for “Professional”, there is evidence that Candidate B is being recommended for their skill and professionalism more than Candidate A is. The implicit assumption here is that longer letters are inherently stronger. If Candidate B's letters are three times longer than Candidate A's, it is highly likely that B will outscore A in most topics for frequency. It is possible for B's letters to feature smaller proportions but higher frequencies than A's in some topics. In some domains, the length of a document may not incapsulate any meaningful information. However, in consulting with letter readers and writers in medicine as well as other academic and professional fields, we concluded simply that if a letter writer spent more time and energy recommending a candidate by writing a longer letter, it could be considered a sign of strength.

Since each topic lexicon was propagated individually, overlap between topics was inevitable. We modified the scoring to address this by dividing the contribution of a token by the number of topic lexicons it appeared in. If, for example, the term “exceptional” appeared in the “Professional”, “Skill”, and “Superlative” topics, its contribution to the scoring of a letter for each of the topics it was in was 1/3. In this way, terms which were specific to a topic, contributed more for the scoring of that topic than words which appeared in many topics.

Section 3.9. Cross-validation

We used cross-validation to test the validity of propagating seed sets derived from real data to measure the topic distribution in new data. To calculate each candidate's scores, we first removed all of the terms in each of topic seed sets that only appeared in the test candidate's letters. For example, if the

term “unequivocally” in the “Superlative” seed set, only appeared in Candidate 6's letters, when testing Candidate 6, “unequivocally” was removed from the “Superlative” seed set before propagating it through WordNet Propagation. However, if the term “best” in the “Superlative” seed set appeared in Candidate 3's and Candidate 6's letters, when testing Candidate 6, “best” was not removed. After removing the test candidate's terms from the seed sets, we then performed and averaged over five folds in which new lexicons were propagated each time. For each fold, a held-out candidate was assigned and the terms in the seed sets which were exclusive to the held-out candidate's letters were removed along with the test candidate's. Since there were six candidates, each candidate's scores were the average over five folds with each of the candidates being the held-out candidate once. Pseudo code for this process is below in Figure 4.

Figure 4. Pseudo code for the method

```
For X in candidates:
    test candidate = X
    For Y in other candidates:
        held-out candidate = Y
        For S in topic seed-sets:
            Remove terms in S exclusive to test candidate and held-out candidate
            Lexicon = WordNetPropagation(S)
            Find test candidate's score for Lexicon
```

We computed the mean and standard deviation for a test candidate's scores across all folds in frequency and proportion for each of the topics. We performed this process for four different iterations through WordNet Propagation.

Section 4: Results

There are a number of angles through which to report and interpret the results. Two scores for each topic – frequency and proportion – were averaged across five folds for each test candidate. This was done for four different iterations of the WordNet Propagation algorithm. Proportion scores serve for within-candidate comparisons and statements about the general make-up of recommendation letters. Frequency scores serve as evidence for comparisons between candidates. The variance in scores through the folds is important for measuring how effective the WordNet Propagation algorithm is at converging on the desired topic lexicons given slight changes in its seed sets. The effect on the scores of different numbers of iterations of the algorithm is also informative. All of the results are available in Appendix D. In this and the next section, only certain aspects of the data are reported and discussed.

Section 4.1. Frequency Scores

Examples of the unnormalized frequency scores are below in Tables 3 and 4 which show the mean and standard deviation for "Skill" and "Teamwork" topics across all candidates and for three iterations of the WordNet Propagation algorithm.

Table 3. Frequency scores for "Skill" topic, WordNet Propagation iterations = 3

"Skill"	Candidate 1	Candidate 2	Candidate 3	Candidate 4	Candidate 5	Candidate 6
μ	48.65	57.91	55.25	52.83	64.35	61.05
σ	1.23	1.28	1.63	1.09	1.18	1.95

Table 4. Frequency scores for "Teamwork" topic, WordNet Propagation iterations = 3

"Teamwork"	Candidate 1	Candidate 2	Candidate 3	Candidate 4	Candidate 5	Candidate 6
μ	35.52	34.69	39.91	30.39	51.21	52.38
σ	0.68	2.72	2.15	0.49	3.28	1.22

The first row corresponds to the average number of hits in each candidate's letters for the "Skill" topic lexicon. The second row is the standard deviation. A correlation exists between the length of a candidate's letters and their frequency scores. However, the relative ranking of frequency scores was

different for each topic despite the effect of letter length, which was constant. The letter lengths are displayed in Table 5 above the rankings for each topic in Table 6.

Table 5. Number of words in concatenated letters for each candidate

	Candidate 1	Candidate 2	Candidate 3	Candidate 4	Candidate 5	Candidate 6
Length	661	772	581	603	699	810

Table 6. Frequency rankings along topics and letter length, iterations = 3

	Candidate 1	Candidate 2	Candidate 3	Candidate 4	Candidate 5	Candidate 6
Skill	6	3	4	5	1	2
Professional	5	3	4	6	2	1
Teamwork	4	5	3	6	2	1
Personal	5	4	6	3	2	1
Compassion	5	3	6	2	1	4
Superlative	4	2	5	6	3	1
Length	4	2	6	5	3	1

Candidate 6, whose letters were the longest, was the highest ranked in four of the six categories. More on the interaction between length and the number of propagation iterations is discussed in <SECTION>. Lack of sufficient data prevented any informative analysis of the true magnitude of the effect. It is possible that despite the correlation, frequency scores do encode meaningful information about the strength of a recommendation, separate from the length of the letters.

Section 4.2. Proportion Scores

In postulating the topics, we hoped to define a set that would account for a large number of tokens in the data used to describe the candidate.[18] The unrealistic goal is to define topic seed sets such that after passing them through WordNet Propagation, the lexicons account for one hundred percent of the content tokens in the data while still preserving the semantic properties of the topics. The

[18] i.e. Not including stopwords, which were removed in preprocessing, or domain-specific stopwords like “medical”, or “obstetrics”.

proportion scores were interesting for evaluating this goal. Tokens in the data which the lexicons did not account for were tallied. The effect of the number of iterations on the proportion of unaccounted for tokens is discussed in <SECTION>.

The pie charts in Figures 5 and 6 below show the topic mixtures, excluding unaccounted for tokens, for Candidates 3 and 6 after three iterations of WordNet Propagation.

Figure 5. Topic mixture for Candidate 3 excluding unaccounted for tokens, iterations = 3

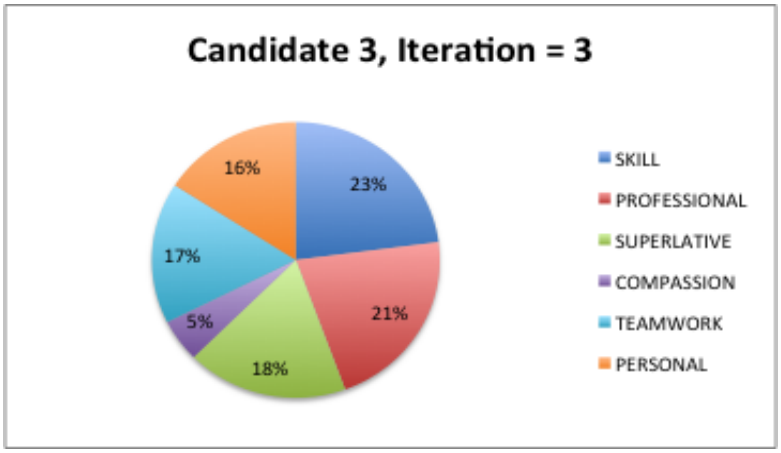
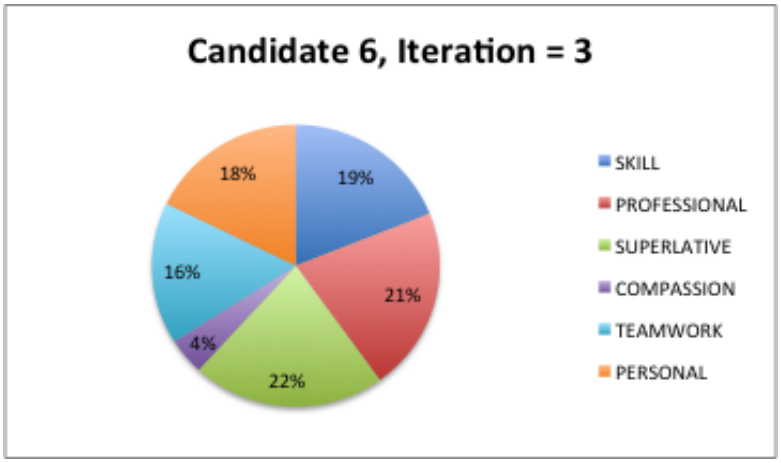
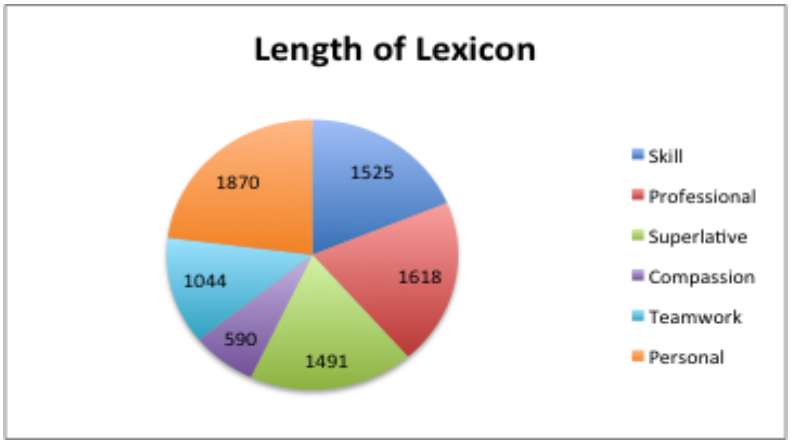


Figure 6. Topic mixture for Candidate 6 excluding unaccounted for tokens, iterations = 3



After three iterations, the topic mixtures are very similar between Candidates 3 and 6. This may reflect some signature make-up of recommendation letters. More likely however, it reflects the various lengths of the lexicons after three iterations. Figure 7 depicts the length of the lexicons in the same pie chart format.

Figure 7. Length of lexicon



The proportion scores were correlated with the length of the lexicons. This is not surprising since they were unnormalized for lexicon length. The ranking of topic proportion across candidates after three iterations is displayed in Table 7 alongside the ranking of lexicon length.

Table 7. Proportion rankings along topic and lexicon length, iteration = 3

	Candidate 1	Candidate 2	Candidate 3	Candidate 4	Candidate 5	Candidate 6	Lexicon Length
Skill	1	1	1	1	1	3	3
Professional	2	2	2	3	2	2	2
Superlative	3	3	3	4	3	1	4
Compassion	6	6	6	6	6	6	6
Teamwork	5	5	4	5	4	5	5
Personal	4	4	5	2	5	4	1

Like for the frequency scores with letter length, it was not possible to measure the true magnitude of the lexicon length effect on proportion scores. The “Compassion” topic, which was the smallest by far, was ranked last for every candidate. However, the effect of lexicon length was clearly not the only thing contributing to the topic mixtures. The “Personal” topic, which was the largest lexicon after three

iterations, was not first ranked for any candidate while the "Skill" topic, which was the third largest lexicon, was the highest ranked for five of the six candidates.

Assuming that the proportion scores were heavily influenced by lexicon length, are the measured topic mixtures still meaningful? If lexicon length were random, then it would be difficult to infer anything from the proportion scores. However the number of words in the initial topic seed sets was not random. Seed sets were developed by appending words found in the data which corresponded to the semantic properties of our postulated topics. Therefore, lexicon lengths reflected our perceived topic distribution in real data. It's possible that the measured topic mixtures were self-fulfilling of our prior beliefs. However, new topic lexicons were propagated from slightly different topic seed sets for each test candidate – the seed sets never included tokens taken from that candidates' letters.

Section 4.3. Proportion of "Other"

After three iterations, the topic lexicons accounted for around one third of the tokens in the letters, excluding stopwords. Figures 8 and 9 below are the same as Figures 5 and 6 above but they include unaccounted for tokens, labeled as "Other"

Figure 8. Topic mixture for Candidate 3 after three iterations, including unaccounted for tokens

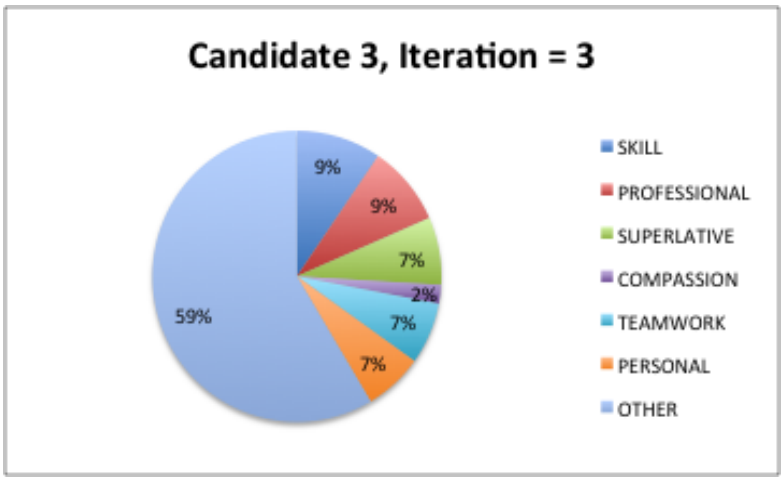
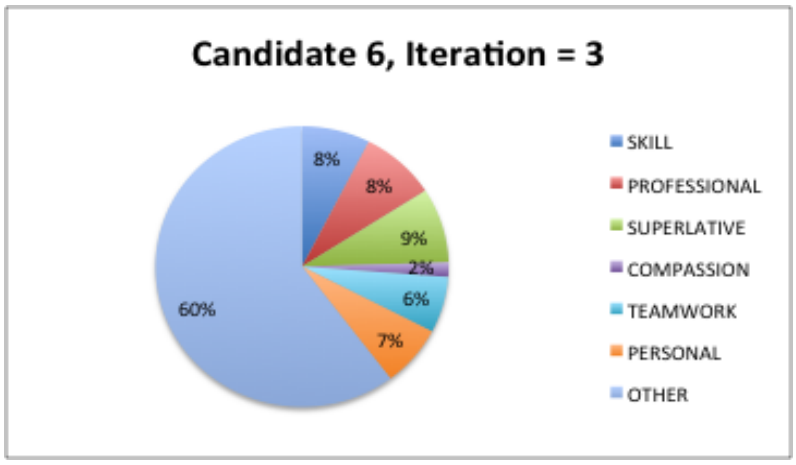


Figure 9. Topic mixture for Candidate 6 after three iterations, including unaccounted for tokens

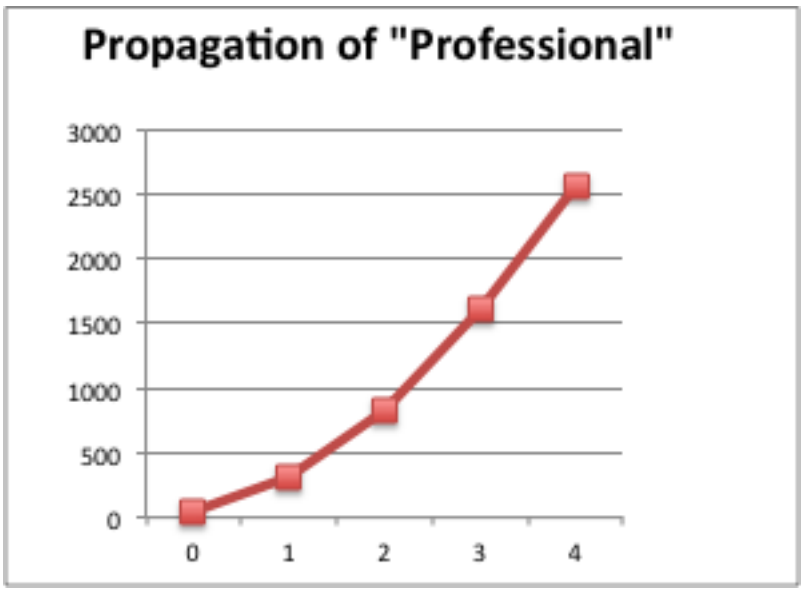


Standard English stopwords were removed during preprocessing. However, domain-specific stop-words like “medicine”, “obstetrics”, or “residency” were not. Therefore, it's difficult to say how successful the topic lexicons were at accounting for the tokens that described the applicant. In the future, a separate domain-specific stop-word list should be developed.

Section 4.4. Number of iterations

The number of iterations for the WordNet Propagation algorithm had far-reaching effects. The most basic effect of the algorithm is that it enlarges the lexicons at a very fast rate. Conceptually, as the number of iterations approaches some large number, the propagated lexicon gets closer to containing all possible words. Below in Figure 10 is the growth rate through four iterations of the number of words in the “Professional” lexicon. The Y-axis is the number of words in the lexicon. The X-axis is the number of iterations.

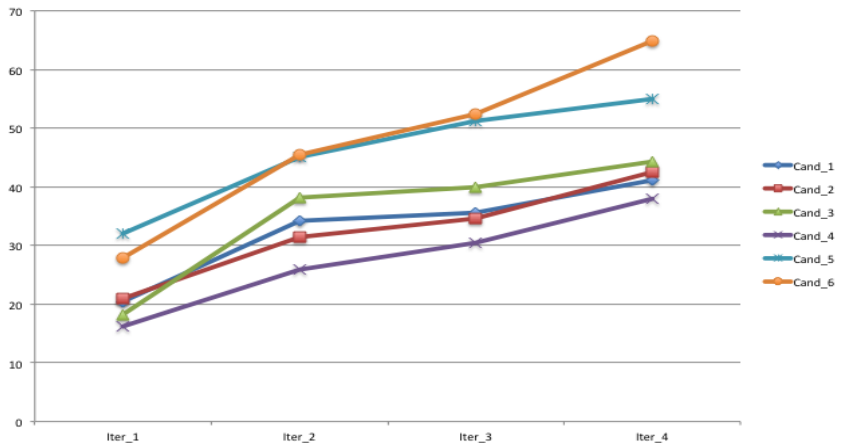
Figure 10. Growth rate of “Professional” lexicon



By the fourth iteration, the “Professional” lexicon already has over 2,500 words, starting with just 41. Many of the terms that appear after four or more iterations do not preserve their topic's semantic properties. However, we expect two things that may mitigate the problems presented by the noise: 1) that all lexicons will suffer a similar proportion of unrelated terms, and 2) that many terms which are unrelated will never appear in the letters and therefore not affect the scoring.

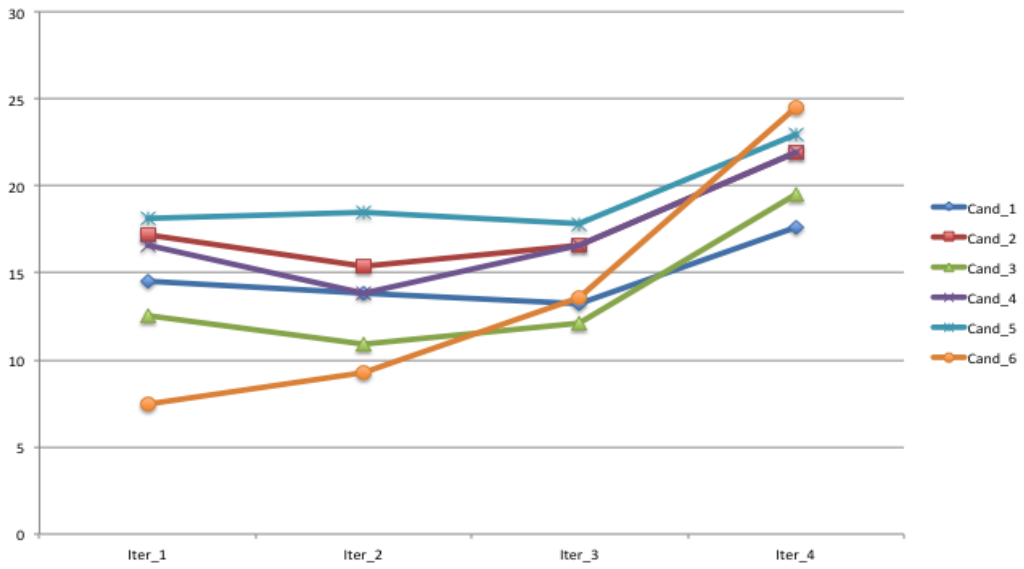
We plotted the frequency scores over the number of iterations for each of the topics. Below in Figure 11 are the scores for the “Teamwork” topic.

Figure 11. Frequency scores for “Teamwork” over number of WordNet Propagation iterations



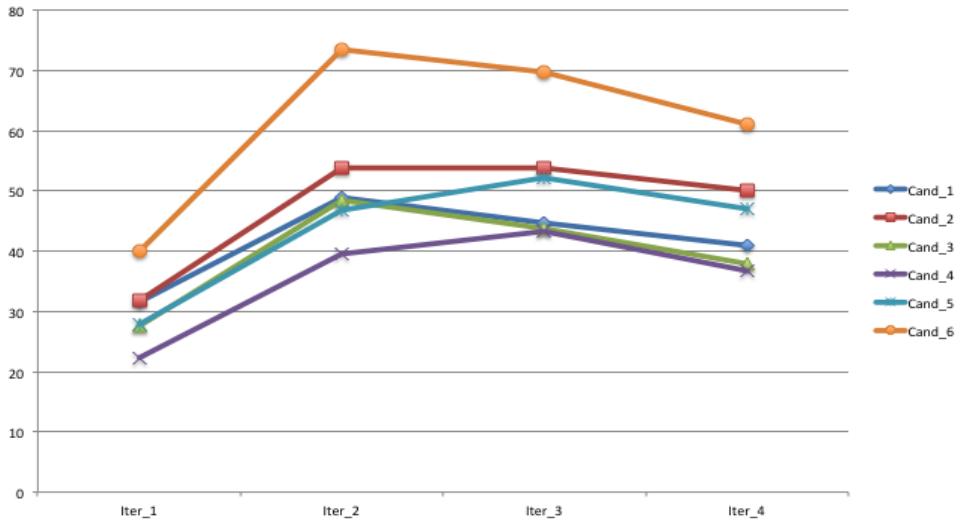
Frequency scores for the “Teamwork” topic increased with the number of iterations. In addition, the relative ranking of candidates changed as the number of iterations increased. Candidate 3, for example, who was fifth ranked after one iteration, was ranked third after four. Changes in ranking could be due to the lexicons becoming more robust. After enough iterations, it is guaranteed that a lexicon contains all words encoding the semantic properties of that topic. It's possible that Candidate 3's letters do contain more words associated with the “Teamwork” topic than the letters for Candidates 2 and 5 and that it simply took a few iterations for the “Teamwork” lexicon to correctly capture that reality. However, it's more likely that the changes in ranking simply reflect the effect of letter length on the frequency scores. As the number of iterations approaches a very large number, the lexicons come closer to containing all possible words. A lexicon which contains all possible words, always give the highest score to the longest letter. Thus, we see that the candidates with the longest letters tended to have the highest scores in every category after enough iterations. In Figure 12 below, the frequency scores for “Compassion” are plotted over the number of iterations. It illustrates this point quite saliently as Candidate 6 (orange), who has the longest letters, begins lowest ranked and ends top ranked by the fourth iteration.

Figure 12. Frequency scores for “Compassion” over number of iterations



In two of the other topics we observed something which at first seemed very strange. The plots of frequency over number of iterations for the "Superlative" and "Skill" topics depict the scores decreasing as the number of iterations increase. If the frequency scores were just tallies of the terms from each topic in the letters, this would be impossible – even if an iteration through the algorithm increased the number of terms in a lexicon by a very small number, the terms from the previous iteration would all still be in that lexicon. Thus, at the very least, the score would remain the same and certainly not decrease. But, the contribution of a token was divided by the number of topic lexicons it appeared in. As the number of iterations increased, the number of topics that included certain terms increased as well, and those tokens' contributions decreased. If the term "best" only appeared in the "Superlative" topic to start, its contribution to the frequency score for "Superlative" began at 1. However, if after 3 iterations the term "best" appeared in the "Superlative", "Personal", and "Skill" topics, its contribution towards the frequency scores for each of those topics became 1/3. It appears that two of the topics – "Superlative" and "Skill" – were redundant: they included enough overlapping terms that their frequency scores decreased as the number of iterations increased. This was particularly true for "Superlative". Figures 12 and 13 illustrate this point.

Figure 13. Frequency scores for "Superlative" over number of iterations



The frequency of unaccounted for tokens predictably declined as the number of iterations increased. This is plotted in Figure 14.

Figure 14. Frequency scores for "Skill" over number of iterations

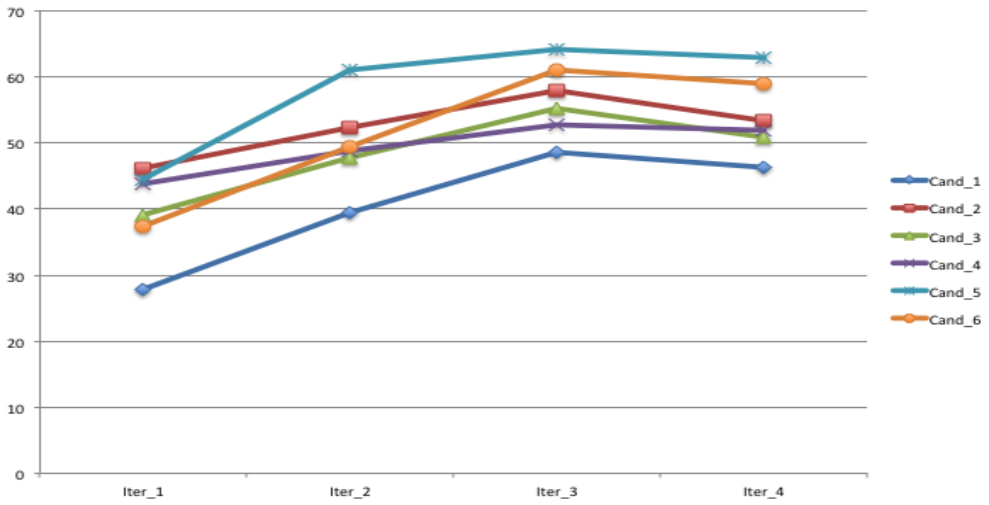
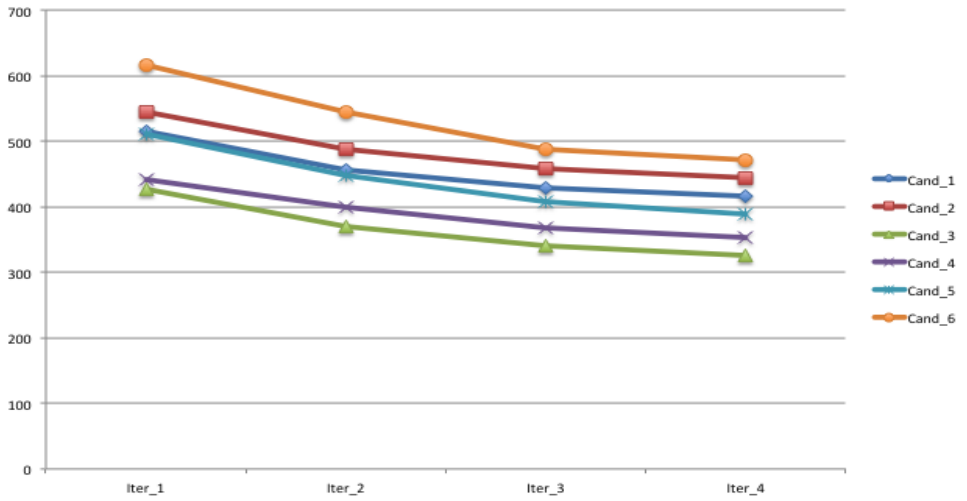
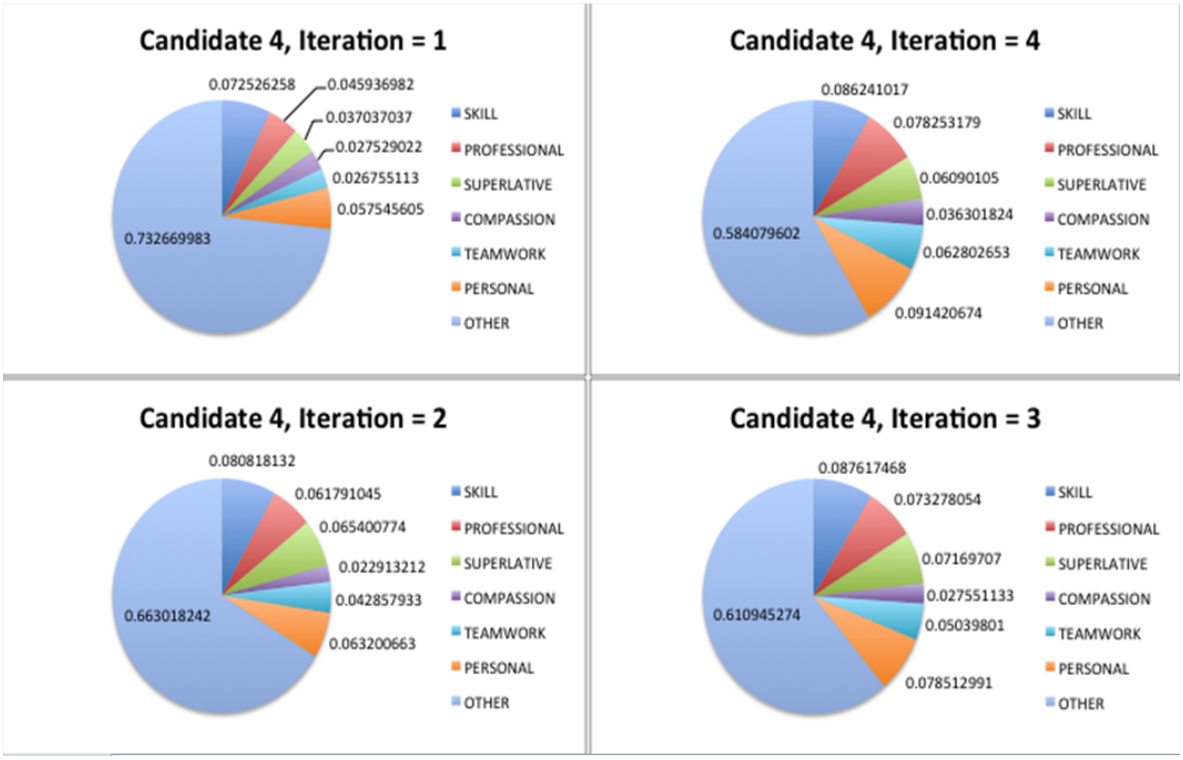


Figure 15. Frequency scores for "Other" over number of iterations



The proportion of unaccounted for tokens across all candidates also decreased as the number of iterations increased. Pie charts for the topic mixture of Candidate 4's letters across all four iterations are below in Figure 15.

Figure 16. Topic mixture for Candidate 4 over four iterations



Section 4.5. Summary

In the end, it was difficult to determine whether the frequency and proportion scores were meaningful because a lack of data prevented an analysis of the magnitude of the effects of letter length and lexicon length. In the future, a separate set of normalized scores should be calculated for comparison.

The fact that certain frequency scores declined as the number of iterations increased was the most important and informative result. It provides valuable feedback for two things: 1) the optimal number of iterations for the algorithm, 2) how to revise the initial proposed topic distribution. The point at which frequency scores across all topics begin to decline is the point at which the lexicons have been propagated too widely and now share all of each others terms. Alternatively, if one or two lexicons feature decreasing frequency scores much earlier than other topics – as we saw with "Superlative" and "Skill – it suggests that these topics are redundant. The ideal set of initial seed sets is one that, after propagation, leads to non-overlapping lexicons which account for a large proportion of

the tokens in the data. When frequency scores decline as iterations increase, the lexicons have significant overlap.

The proportion of unaccounted for tokens is also an important measure of the algorithm's behavior. In the future, a set of domain-specific stopwords should be developed. Terms like “medical” or “obstetrics” are not descriptive of the candidate. By removing these along with the standard stopwords (e.g. “and”, “the”), the proportion of unaccounted for tokens will more closely reflect the proportion of content tokens not captured by the method. This measure, in conjunction with measures about the rate of change of the frequency scores as the number of iterations increase, would be very useful in postulating a revised mixture over topics.

Section 5: Conclusion

Much of the sentiment analysis literature is geared towards developing systems which can autonomously classify and organize opinion in text. The first requirement of these systems is a large volume of annotated data on which to train. Rich annotated data does not always exist in abundance. Moreover, autonomous systems may not be very useful or desirable in certain domains. The recommendation letters project is an example this. Will the admissions board of a residency program ever cede the process of careful vetting to a classifier?

The method outlined here is intended for use by domain experts to explore and organize their intuitions. It allows experts to test their beliefs about the topical make-up of the documents off of which they make critical decisions. The initial goal is to define a set of topic seed sets which, after a proper number of iterations, propagate into non-overlapping topic lexicons which preserve the original semantic properties and account for a large portion of the substantive tokens in the documents. Regardless of if an expert ever achieves this goal, the process informs their beliefs. In applying this method to recommendation letters, we learned that the “Superlative” and “Skill” topics we postulated were redundant after only one iteration. In other words, the signal they were capturing was already

being captured by the other topics. We also learned that our topics only accounted for one third of the content tokens in the data. With this knowledge, we could re-read the letters, re-think our beliefs about their topical make-up, and try again. By proceeding in this way, we unpack the blackbox of our own intuition which can be simultaneously powerful and inconsistent. We can arrive at a better understanding of our qualitative beliefs through organized processes like the one outlined here.

Section 6: Future Directions

This project is ongoing, and as we resolve obstacles to compiling the dataset, a number of promising avenues open up. In this last section, a few of the data-driven directions we've discussed are outlined.

Section 6.1. Linear Models

Constructing a series of linear models with linguistic features from the text and associated metadata would be the first step in analyzing a large dataset. An application, which includes three reference letters and the Dean's Letter, comes with the applicant's GPA, MCAT scores, and class ranking. Constructing different linear models with each of those metrics as the dependent variable may yield quick and potentially valuable insights. Linguistic features could be, for example, the proportion of superlative to non-superlative adjectives, the length of the letter, the topic-lexicon scores developed here. They could also be more sophisticated. One idea is to calculate the proportion of past to present tense verbs in a letter. The intuition behind this feature is that accomplishments are reported in the past tense while vaguer personal characteristics are reported in the present:

1) "She cracked one of the hardest problems on my exam."
2) "He is very bright and eager to learn."

In general, statements that cannot be made about everyone seem stronger. While a letter-writer can truthfully assert that everyone is "bright and eager" they cannot credit applicants with accomplishments they did not accomplish.

As previously mentioned, a system which accurately predicts one of the provided quantitative metrics based on linguistic features in the text could be used to flag mismatches. The admissions board would be well-served to know that several applicants they skipped over due to below-average GPAs or MCAT scores have letters written about them as if they had high scores.

Another route more exciting route involves two qualitative scores. One feature which could be developed would be the mean human judgment of a letter. If a handful of admissions staff read each letter and rated it on a five-point scale, we could train a classifier to predict the mean human judgment based on linguistic features and associated metadata. Alternatively, a dataset could be compiled from only letters written for applicants that went through the residency program and residency staff members could rate their performance on a five-point scale. A linear model with this as its dependent variable could yield very interesting results. Such a system could discover topics and features in the letters that predict an applicant's success that were previously unknown to human reviewers. These two approaches – predicting the human judgment score and predicting an applicant's success score – represent two fundamentally different systems. The goal of the first system is to automate the current process while the goal of the second is to inform it – both could play an important role.

Section 6.2. Latent Dirichlet Allocation

The method outlined in this paper, through which a domain expert explores their intuition about the topical make-up of a document, is conceptually analogous to inference in Latent Dirichlet Allocation (LDA). Given a set of documents, the expert asserts the most likely mixture over topics. The generative process of LDA is also very attractive for modeling the generative story of a recommendation letter. A letter writer has limited space to telegraph a small handful of positive characteristics about the applicant. Each word takes up valuable space, so there is an incentive for the author to make sure that every word is helping to telegraph at least one or more of these characteristics. In this way, the process through which an author generates a letter is by choosing a characteristic to

telegraph, choosing a word to do so, and repeating until the author runs out of space. This is LDA's generative process: choose a topic from a distribution over topics, choose a word from a distribution over words. Similarly, the process of reading a letter and inferring its intended message is analogous to inference: for each word, infer its most likely topic assignment. The original research question that was posed by the doctors at UMass Worcester was essentially this. They asked, is there a systematic way to resolve disputes over the true meaning of words and phrases in the letters. In other words, which topics generate which words?

The output of LDA could be used descriptively in pursuing the question of topical make-up. It could also be used to produce certain features to add into a linear model. For example, with a large enough dataset, we could calculate a mean topic distribution for letters of recommendation and for a new incoming letter, calculate its distance. This seems like an important feature – a measure of uniqueness, calculated using some distance measure like KL-divergence. This idea could be further enhanced by binning the dataset. Instead of calculating the mean topic distribution across all letters, we could, for example, calculate means for letters corresponding to applicants within certain GPA ranges.

Section 6.3. Strategy for data collection

The recommendation letters project will be as successful as it is able to compile a large, systematized dataset. For a number of reasons, this is unusually challenging. Issues of privacy are the foremost hurdle in rapidly acquiring data in digital form. Moreover, the most promising directions are those that require a qualitative scoring of applicants and their letters. Compiling these scores will require the time and effort of doctors and medical professionals who have little, if any, to spare.

But the benefit of such a system should be readily apparent even if only partially developed. The project should aim to gain concrete results from limited data to demo its potential to other residency programs. The more data this project attracts, the more insightful and valuable it will become – it has the potential to snowball in this way and lead to fascinating insights about the admissions process and advance sentiment analysis as a field.

Appendix A: Sample Data:

A.1. Letter of recommendation, Candidate 3

September 30, 2010

Dear Program Director,

It is our pleasure to write a letter of recommendation for xxAPPLICANTxx for an Ob/Gyn residency. We have known xxAPPLICANTxx since she was a student in the Ob/Gyn clinical clerkship at Boston Medical Center. She has just begun her sub-internship in gynecology, and has taken on a number of organizational tasks for the group of students going into Ob/Gyn this year. xxAPPLICANTxx during her 3rd year rotation stood out particularly for her compassion and advocacy for her patients. She worked with a challenging group of patients (substance abuse in pregnancy) and was able to forge excellent relationships with the care team and the patients to identify their needs and help direct them to resources. xxAPPLICANTxx is also a successful researcher; publishing papers during her laboratory research time, as well as working with our Family Planning research team at BU to present their work at the Reproductive Health Conference in 2009. As you can see from her personal statement her impetus to patient care is driven by her strong family influences, which have led her to Ob/Gyn. xxAPPLICANTxx is a calm, detail focused hard worker as well as a strong team member who is able to quickly rally resources to advocate for her patients.

xxAPPLICANTxx is also a great teacher of medical students and her patients. Because of her strong communication skills she is able both identify when people are overwhelmed, and is able to stop in the moment to break down a situation into an appropriate discussion at each person's level of need. The residents and staff found that because she was so reliable, they could delegate activities to her and not worry that anything would be missed. She has been able to truly function as an outstanding acting intern in her rotations on our service.

xxAPPLICANTxx is one of ten students from Boston University School of Medicine who are applying for residency training in obstetrics and gynecology this year. Our "30,000 foot" view of her performance and attributes place her as a very strong candidate.

xxAPPLICANTxx has waived her right to see this letter under the "Family Educational Rights and Privacy Act". Please do not hesitate to contact either of us if we can be of any further assistance. Sincerely, xxAUTHORxx

A.2. Letter of recommendation, Candidate 6

September 30, 2010

Dear Program Director,

It is our pleasure to write a letter of recommendation on behalf of xxAPPLICANTxx for a residency in obstetrics and gynecology. XxAPPLICANTxx did her third year clerkship in obstetrics and gynecology and a sub-internship in gynecology with us. She began working with our patients at Boston Medical Center prior to medical school as a volunteer in the Grow Clinic, a multi-discplinary pediatric program that includes food and legal access. XxAPPLICANTxx herself has a compelling personal story, which has clearly shaped her views on healthcare, education and poverty. What makes her an outstanding future physician is the way in which her experiences have shaped the choices she has made in preparing herself for her future career.

xxAPPLICANTxx due to both her family background and her experiences in the Grow clinic, felt that a standard medical school curriculum would not give her the tools she desired to understand and incorporate women's health, poverty and policy into her future practice. So she decided to earn a public health degree as part of her medical education. Working with classmates and faculty, she designed an advocacy curriculum to supplement the traditional course path for medical students here at BUSM; aspects of this program will begin during this academic year.

During both her clerkship and her sub-internship with us, xxAPPLICANTxx was able to demonstrate her compassion as well as her focused skills in patient care. Although she was concerned about doing her sub-internship immediately upon completion of her MPH, she jumped right in and truly performed at the level of an intern. She was on service when our new interns started and she both oriented them to the systems of care in our institution and helped run the service.

As impressive as it is, none of the above information actually gives you a feel for what a strong, determined individual she is, missing as well is her warmth and humble accounting of her accomplishments. Undoubtedly she achieve her goals of improving health care for women and their families, in addition to being a tremendous asset to any training program wise enough to match her.

xxAPPLICANTxx is one of ten students from Boston University School of Medicine who are applying for residency training in obstetrics and gynecology this year. Our “30,000 foot” view of her performance and attributes place her as one of our top two candidates.

xxAPPLICANTxx has waived her right to see this letter under the “Family Educational Rights and Privacy Act”. Please do not hesitate to contact either of us if we can be of any further assistance.

Sincerely, xxAUTHORxx

Appendix B: WordNet Propagation Algorithm

WordNetSensePropagate($S, iter$)

▷ Input
▷ S: a list of synsets. For example: $\langle$ {brilliant·s·01, n·win·01}, {sadly·r·01, gross·a·01} $\rangle$
▷ $iter$: the number of iterations
▷ Output
▷ T: Length(S) × 1+$iter$ synset matrix: $\begin{pmatrix} S[1] & \cdots & iter\text{-th propagation of } S[1] \\ \vdots & & \vdots \\ S[\text{Length}(S)] & \cdots & iter\text{-th propagation of } S[n] \end{pmatrix}$

initialize T: a Length(S) × 1+$iter$ matrix such that $T[i][1] = S[i]$ for $1 \leqslant i \leqslant 1+iter$
for $i \leftarrow 1$ **to** $iter$
    for $j \leftarrow 1$ **to** Length(S)
        $newSame \leftarrow$ SamePolarity($T[j][i]$)
        $others \leftarrow \bigcup_{k=1}^{\text{Length}(S)} T[k][i]$ for $k \neq j$ ▷ The other seed-sets in this column.
        $newDiff \leftarrow$ OtherPolarity($others$)
        ▷ For the experiments, I first calculate all the propagation sets and then eliminate their
        ▷ pairwise intersection from each, to ensure no overlap.
        $T[j][i+1] \leftarrow (newSame \cup newDiff)$
return T

SamePolarity($synsets$)

$newsynsets \leftarrow \{\ \}$
for $s \in synsets$ ▷ Synset-level relations.
    $newsynsets \leftarrow newsynsets \cup \{s\} \cup$ **AlsoSees**(s) ∪ **SimilarTos**(s)
    for $lemma \in$ **Lemmas**(s) ▷ Lemma-level relations.
        for $altLemma \in$ (**DerivationallyRelatedForms**($lemma$) ∪ **Pertainyms**($lemma$))
            $newsynsets \leftarrow newsynsets \cup \{$**Synset**($altLemma$)$\}$
return $newsynsets$

OtherPolarity($synsets$)

$newsynsets \leftarrow \{\ \}$
for $s \in synsets$
for $lemma \in$ **Lemmas**(s) ▷ Lemma-level relations.
    for $altLemma \in$ **Antonyms**($lemma$)
        $newsynsets \leftarrow newsynsets \cup \{$**Synset**($altLemma$)$\}$
return $newsynsets$

Appendix C: Seed sets

C.1. "Skill"

"Skill topic seed set = [skill, intelligence, insightful, established, asset, valued, challenging, challenge, experience, grasp, competent, knowledge, essential, keen, investigatory, talented, ability, innate, achievement, achieving, handily, clinical, acumen, strong, specialty, research, performance, judgment, intellect, aptitude, understanding, intuition, dexterity, capacity, foundation, background, award, improve, ability, successfully, wise, excel, analytical]

C.2. "Professional"

"Professional" topic seed set = [professional, responsible, aplomb, serious, discipline, commitment, prepared, maturity, mature, solid, functioned, efficient, exhaustive, attention, dedicated, volition, motivated, pursue, focus, management, thorough, task, independently, undaunted, driven, devotion, diagnostic, systematic, determination, goal, career, timely, detail, proactive, busy, organized, resilient, positive, systematic, straightforward, determination]

C.3. "Personal"

"Personal" topic seed set = [warm, thoughtful, engaging, respectful, sincere, friendly, pleasant, helpful, enthusiastic, pleasure, active, ethical, easy-going, manner, care, eager, loves, passion, well-rounded, well rounded, diverse, sensitive, polite, bright, actively, energy, confident, respect, dignity, energetic, delightful, optimistic, cheery, passionate, comfortable, positive, warm , sensitive, humor, delightful]

C.4. "Compassion"

"Compassion" topic seed set = [compassionate, global, refugee, patients, care, hunger, international, advocacy, conscientious, bedside, social, ethnic, economic. humanism, community, youth, communities, legal, food, poverty, overcome]

C.5. “Teamwork”

“Teamwork” topic seed set = [involved, communication, together, team, concise, clear, informative, participated, expressed, interacted, members, trusting, relationships, rapport, feedback, reliably, interpersonal, leader, organizational, leadership, contribution, engagement, entrusted, articulates, extracurricular, volunteered, network, advice, support, listener, collaborative]

C.6. “Superlative”

“Superlative” topic seed set = [unsurpassed, unprecedented, unfailingly, unequivocally, top, best, consummate, exceptional, excellent, excel, optimally, unusually, exceedingly, lauded, terrific, extraordinary, beyond, perfect, utmost, special, superb, wonderful, beyond, brightest, most, very, strongest]

Appendix D: Results

D.1. Frequency scores

D.1.1. WordNet Propagation iterations = 1

Cand_1	SKILL	PROF	SUP	COMP	OTHER	TEAM	PERS
μ	27.73	31.00	31.57	14.53	515.20	20.43	20.53
σ	0.72	0.94	2.77	1.19	2.59	0.75	1.25
Cand_2							
μ	46.19	32.02	31.72	17.20	544.80	20.87	29.20
σ	0.79	2.33	2.22	1.79	4.44	2.77	1.49
Cand_3							
μ	39.12	35.95	27.58	12.57	427.20	18.23	20.35
σ	0.62	1.20	2.70	0.89	4.02	1.11	1.69
Cand_4							
μ	43.73	27.70	22.33	16.60	441.80	16.13	34.70
σ	1.39	1.28	5.63	0.22	5.36	0.51	2.82
Cand_5							
μ	44.51	31.91	27.88	18.13	510.20	31.98	34.39
σ	2.70	1.91	4.43	1.57	5.36	0.98	4.66
Cand_6							
μ	37.37	49.13	40.00	7.43	616.40	27.80	31.87
σ	1.42	1.59	5.74	0.37	6.99	1.56	2.05

D.1.2. WordNet Propagation iterations = 2

Cand_1	SKILL	PROF	SUP	COMP	OTHER	TEAM	PERS
μ	39.46	41.29	48.86	13.81	456.00	34.25	27.34
σ	2.41	1.84	2.04	1.07	2.35	0.79	2.52
Cand_2							
μ	52.38	44.91	53.83	15.38	487.60	31.43	36.48
σ	1.95	3.36	2.60	1.74	1.52	2.72	2.04
Cand_3							
μ	47.79	39.17	48.45	10.89	370.60	38.20	25.89
σ	3.47	0.91	2.07	1.15	4.93	1.58	1.90
Cand_4							
μ	48.73	37.26	39.44	13.82	399.80	25.84	38.11
σ	2.64	1.75	11.08	0.34	8.11	0.41	2.82
Cand_5							
μ	60.97	44.96	46.71	18.45	448.20	45.02	34.68
σ	3.67	2.37	8.97	1.73	7.40	2.56	3.93
Cand_6							
μ	49.53	51.97	73.41	9.25	543.60	45.58	36.66
σ	3.36	1.58	3.09	0.84	3.21	1.05	1.72

D.1.3. WordNet Propagation iterations = 3

Cand_1	SKILL	PROF	SUP	COMP	OTHER	TEAM	PERS
μ	48.65	48.09	44.71	13.19	429.20	35.52	41.64
σ	1.23	4.03	2.70	0.54	4.15	0.68	5.74
Cand_2							
μ	57.91	57.14	53.76	16.56	457.40	34.69	44.53
σ	1.28	6.42	3.64	1.62	3.65	2.72	2.90
Cand_3							
μ	55.25	51.82	43.66	12.07	340.00	39.91	38.28
σ	1.63	2.53	2.97	0.76	3.67	2.15	0.59
Cand_4							
μ	52.83	44.19	43.23	16.61	368.40	30.39	47.34
σ	1.09	1.13	7.81	1.35	5.59	0.49	1.93
Cand_5							
μ	64.25	57.59	52.21	17.83	407.20	51.21	48.71
σ	1.18	2.69	4.99	1.63	6.83	3.28	4.09
Cand_6							
μ	61.05	67.85	69.73	13.53	488.40	52.38	57.05
σ	1.95	3.99	5.00	0.76	0.89	1.22	0.98

D.1.4. WordNet Propagation iterations = 4

Cand_1	SKILL	PROF	SUP	COMP	OTHER	TEAM	PERS
μ	46.28	49.28	41.05	17.62	41.15	49.62	416.00
σ	1.88	2.52	1.96	0.76	1.33	6.28	4.53
Cand_2							
μ	53.37	59.58	50.02	21.88	42.51	50.43	444.20
σ	1.25	6.95	3.61	0.83	2.19	4.20	3.70
Cand_3							
μ	50.82	54.35	37.98	19.54	44.23	49.08	325.00
σ	1.43	1.74	2.14	0.96	1.90	1.44	1.22
Cand_4							
μ	52.00	47.19	36.72	21.89	37.87	55.13	352.20
σ	1.24	4.02	5.51	1.23	0.61	2.97	1.79
Cand_5							
μ	62.91	63.86	46.92	22.91	54.93	58.88	388.60
σ	1.13	3.13	3.74	1.71	1.60	4.25	4.67
Cand_6							
μ	59.06	68.13	61.02	24.47	64.80	61.13	471.40
σ	2.17	2.24	2.01	2.18	1.17	1.34	1.14

D.2. Proportion scores

D.2.1. WordNet Propagation iterations = 1

Cand_1	SKILL	PRO	SUP	COMP	TEAM	PERS	OTHER
μ	0.0420	0.0469	0.0478	0.0220	0.0309	0.0311	0.7794
σ	0.0011	0.0014	0.0042	0.0018	0.0011	0.0019	0.0039
Cand_2							
μ	0.0640	0.0443	0.0439	0.0238	0.0289	0.0404	0.7546
σ	0.0011	0.0032	0.0031	0.0025	0.0038	0.0021	0.0061
Cand_3							
μ	0.0673	0.0619	0.0475	0.0216	0.0314	0.0350	0.7353
σ	0.0011	0.0021	0.0047	0.0015	0.0019	0.0029	0.0069
Cand_4							
μ	0.0725	0.0459	0.0370	0.0275	0.0268	0.0575	0.7327
σ	0.0023	0.0021	0.0093	0.0004	0.0008	0.0047	0.0089
Cand_5							
μ	0.0637	0.0457	0.0399	0.0259	0.0457	0.0492	0.7299
σ	0.0039	0.0027	0.0063	0.0022	0.0014	0.0067	0.0077
Cand_6							
μ	0.0461	0.0607	0.0494	0.0092	0.0343	0.0393	0.7610
σ	0.0018	0.0020	0.0071	0.0005	0.0019	0.0025	0.0086

D.2.2. WordNet Propagation iterations = 2

Cand_1	SKILL	PRO	SUP	COMP	TEAM	PERS	OTHER
μ	0.0597	0.0625	0.0739	0.0209	0.0518	0.0414	0.6899
σ	0.0036	0.0028	0.0031	0.0016	0.0012	0.0038	0.0035
Cand_2							
μ	0.0725	0.0622	0.0746	0.0213	0.0435	0.0505	0.6753
σ	0.0027	0.0047	0.0036	0.0024	0.0038	0.0028	0.0021
Cand_3							
μ	0.0823	0.0674	0.0834	0.0187	0.0658	0.0446	0.6379
σ	0.0060	0.0016	0.0036	0.0020	0.0027	0.0033	0.0085
Cand_4							
μ	0.0808	0.0618	0.0654	0.0229	0.0429	0.0632	0.6630
σ	0.0044	0.0029	0.0184	0.0006	0.0007	0.0047	0.0134
Cand_5							
μ	0.0872	0.0643	0.0668	0.0264	0.0644	0.0496	0.6412
σ	0.0052	0.0034	0.0128	0.0025	0.0037	0.0056	0.0106
Cand_6							
μ	0.0611	0.0642	0.0906	0.0114	0.0563	0.0453	0.6711
σ	0.0041	0.0019	0.0038	0.0010	0.0013	0.0021	0.0021

D.2.3. WordNet Propagation iterations = 3

Cand_1	SKILL	PRO	SUP	COMP	TEAM	PERS	OTHER
μ	0.0736	0.0728	0.0676	0.0200	0.0537	0.0630	0.6493
σ	0.0019	0.0061	0.0041	0.0008	0.0010	0.0087	0.0063
Cand_2							
μ	0.0802	0.0791	0.0745	0.0229	0.0481	0.0617	0.6335
σ	0.0018	0.0089	0.0050	0.0022	0.0038	0.0040	0.0051
Cand_3							
μ	0.0951	0.0892	0.0751	0.0208	0.0687	0.0659	0.5852
σ	0.0028	0.0044	0.0051	0.0013	0.0037	0.0010	0.0063
Cand_4							
μ	0.0876	0.0733	0.0717	0.0276	0.0504	0.0785	0.6109
σ	0.0018	0.0019	0.0130	0.0022	0.0008	0.0032	0.0093
Cand_5							
μ	0.0919	0.0824	0.0747	0.0255	0.0733	0.0697	0.5825
σ	0.0017	0.0038	0.0071	0.0023	0.0047	0.0058	0.0098
Cand_6							
μ	0.0754	0.0838	0.0861	0.0167	0.0647	0.0704	0.6030
σ	0.0024	0.0049	0.0062	0.0009	0.0015	0.0012	0.0011

D.2.4. WordNet Propagation iterations = 4

Cand_1	SKILL	PRO	SUP	COMP	TEAM	PERS	OTHER
μ	0.0700	0.0746	0.0621	0.0267	0.0622	0.0751	0.6293
σ	0.0028	0.0038	0.0030	0.0012	0.0020	0.0095	0.0068
Cand_2							
μ	0.0739	0.0825	0.0693	0.0303	0.0589	0.0699	0.6152
σ	0.0017	0.0096	0.0050	0.0012	0.0030	0.0058	0.0051
Cand_3							
μ	0.0875	0.0936	0.0654	0.0336	0.0761	0.0845	0.5594
σ	0.0025	0.0030	0.0037	0.0017	0.0033	0.0025	0.0021
Cand_4							
μ	0.0862	0.0783	0.0609	0.0363	0.0628	0.0914	0.5841
σ	0.0021	0.0067	0.0091	0.0020	0.0010	0.0049	0.0030
Cand_5							
μ	0.0900	0.0914	0.0671	0.0328	0.0786	0.0842	0.5559
σ	0.0016	0.0045	0.0054	0.0025	0.0023	0.0061	0.0067
Cand_6							
μ	0.0729	0.0841	0.0753	0.0302	0.0800	0.0755	0.5820
σ	0.0027	0.0028	0.0025	0.0027	0.0014	0.0017	0.0014

BIBLIOGRAPHY

PAPERS

1. J. M. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, “Learning subjective language,” Computational Linguistics, vol. 30, pp. 277–308, September 2004.
2. Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
3. E. Rilo and J. Wiebe, “Learning extraction patterns for subjective expressions,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.
4. J. M. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, “Learning subjective language,” Computational Linguistics, vol. 30, pp. 277–308, September 2004.
5. C. O. Alm, D. Roth, and R. Sproat, “Emotions from text: Machine learning for text-based emotion prediction,” in Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005.
6. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86, 2002

BOOKS

1. “Opinion Mining and Sentiment Analysis”, 2008, B. Pang, L. Lee.
2. “Sentiment Analysis and Opinion Mining”, 2012, B. Liu.

WEBSITES

1. http://sentiment.christopherpotts.net/
2. http://wordnet.princeton.edu/

TALKS

1. C. Potts, “Sentiment Symposium Tutorial,” in Sentiment Analysis Symposium, San Francisco, 2011.
2. B. Liu, “Sentiment Symposium Tutorial”, in Sentiment Analysis Symposium, New York, 2012.