Topic Modeling & Similarity Analysis of Syracuse University Commencement Speeches

Ryan H. French & William D. Lombardi

Syracuse University

rhfrench@syr.edu, wlombard@syr.edu

Introduction

Every Spring another class of graduating Syracuse University Students approaches the dome in caps and gowns prepared to graduate and enter the workforce. The graduation itself is a grand ceremony involving processions, speeches, and the distribution of diplomas. However, the commencement speech tends to be the defining characteristic of each graduation and is our topic of research for today.

Usually selected for success in their field, commencement speakers are meant to congratulate students and welcome them into the professional realm as their new lives “commence”. Over the past decade and a half, Syracuse University has welcomed notable commencement speakers such as former Vice President Joe Biden, Billy Joel, and Rudolph Giuliani, as well as a number of other politicians, businessmen, and artists who have left their mark on the world.

If one were to view some of these commencement speeches, a few clear topics and themes are likely to become apparent involving success, work, and relationships. In order to better understand these trends and determine what elements were generally consistent in a commencement speech, we decided to conduct topic modeling and theme analysis on a corpus of Syracuse University commencement speeches.

Our Objective

In this paper, we will be examining the speeches given by Syracuse University commencement speakers as a collective in order to determine what the most significant overall topics are and further determine which are the most relevant by associating them with the corresponding number of documents per topic. We will then seek to compare these papers individually by examining their cosine similarity and determining how closely they resemble other documents within the corpus. Finally, we will explore the topics themselves and see what they appear to say about the subjects and themes of the documents to best understand the archetypal commencement speech.

Data Collection

Perhaps the most time consuming and difficult phase of this process, data collection proved a significant challenge. Although we initially believed that we had found all of the commencement speeches from 2002 to 2018 archived at commencement.syr.edu, we discovered that a few of these were lacking reliable transcripts or were entirely absent when we attempted to view links for specific speeches.

We attempted to rectify this by both reaching out to the site curators of commencement.syr.edu and the Syracuse University archives, but were unable to find the documents in either of these entities.

While a few speeches still eluded our grasp (2017, 2018, 2009, 2008), we were able to find a number of the previously missing ones at Syracuse University News website resulting in a corpus of 13 total speeches to analyze.

Our Dataset

After locating each of the speeches online, we created a plain text .txt file for each which was stored locally. Speeches range in length from a paragraph to pages, leading to a certain degree of inconsistency in this regard.

Data Cleaning and Preprocessing

When moving documents from the sites they were hosted on to local text files, we encountered a number of strange anomalies. While some of these were rectified by formatting the document as plain text, there were still a few instances where Chinese characters such as 腺 or 鮮 appeared seemingly at random. These occurrences were fixed where the intended word could be made out or removed when it could not.

As part of our tokenization process we removed English stop words (as well as the artifacts 've' and 'don' due to their presence as salient tokens in our topics), converted all tokens to lowercase, and considered both unigram and bigram phrases. Using gensim's simple pre-process, we also removed any punctuation.

After consulting **Machine Learning Plus**, we elected to utilize lemmatization to stem our words while simultaneously adding post tags for part of speech to further increase the accuracy of our modeling (LDA - How to Grid Search Best Topic Models?).

Latent Dirichlet Allocation (LDA) Modeling

In order to perform our topic modeling, we elected to utilize the Latent Dirichlet Allocation (LDA) Modeling available through sklearn.

As mentioned in the pre-processing step, we compared solely unigram models to both unigram and bi-gram models and elected to utilize the ladder as it added an extra element of depth and helped to expand the number of overarching topics present.

We utilized 10 for the maximum number of iterations, 1000 samples, and a random state of 0 for reproducibility purposes.

After doing some research as to mode optimization, we utilized sklearn's Grid Search in order to determine our most effective model. Iterating over the parameters of n_components with a range of 3-6 and learning decay with options of 0.2, 0.3, 0.5, 0.7, and 0.9, the grid search determined that 3 topics with a learning decay of 0.3 would produce the best results.

Our Results

The result of modeling was three topics for our corpus with varying degrees of documents attributed to each. The top 15 tokens for each topic in descending order, as well as the name we attributed to each, are as follows:

	New Beginnings	**Timely Action**	**Worldliness**
1	new	say	people
2	year	year	world
3	change	time	think
4	life	good	life
5	synthetic	today	say
6	cell	know	good
7	people	make	time
8	know	look	know
9	future	day	day
10	energy	life	year
11	genomic	think	just
12	computer	thing	hope

13	dna	people	make
14	genome	way	work
15	make	just	come

Figure 1 - Top 15 Tokens per Topic

The first topic was dubbed New Beginnings due to its emphasis on life, transformation, and genetics. This is exhibited through tokens such as ‘dna’, ‘life’, and ‘change’. The influence of beginning anew makes logical sense for commencement speeches as the intention of a commencement speech is generally to wish students well in their new lives.

Our second topic was titled Timely Opportunity due to its emphasis on time related matters and words that indicate action. Some examples of time based tokens are ‘year’ and time while actionable consideration tokens are ‘think’ and ‘say’. The emphasis on time and action is interesting as it implies the discussion of taking action within certain time periods.

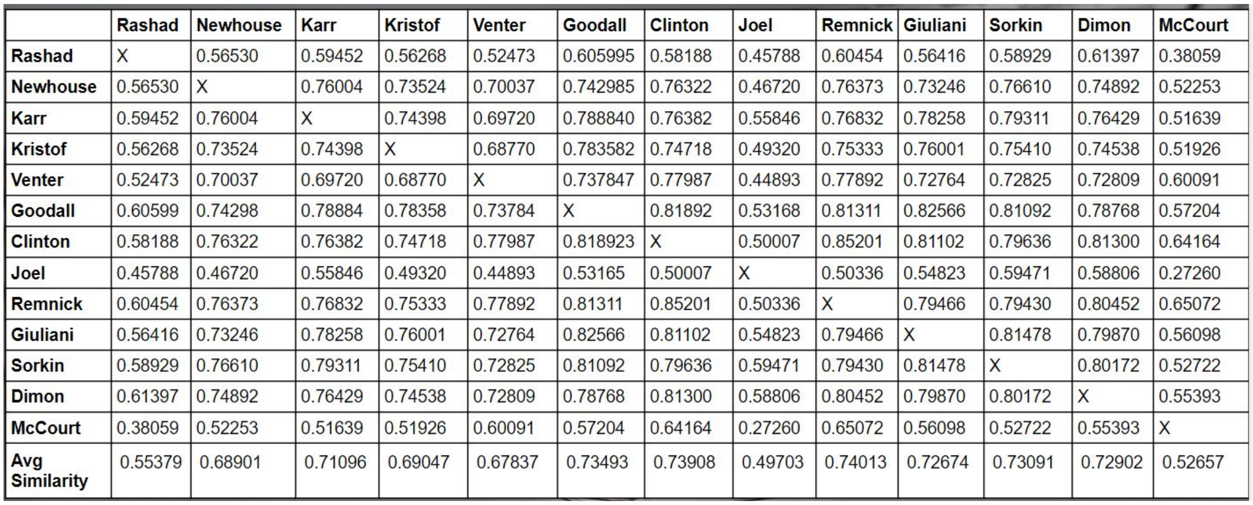
The final topic named Worldliness overlaps heavily with our second topic of Timely Opportunity, which we believe is likely due to the relatively high degree of similarity between the various documents in our corpus. However, a few key differences in tokens and saliency seem to suggest a slightly different meaning focused more on community and a sense of worldliness. This is indicated by the 2 most salient tokens ‘people’ and ‘world’.

Topic	**Number of Documents**
New Beginnings	4
Timely Action	4
Worldliness	5

Figure 2 - Document Cosine Similarity Matrix

Additionally, we found that the document distribution per speech was relatively even, each topic containing approximately a third of the collective corpus of documents.

Document Similarity Analysis



	Rashad	Newhouse	Karr	Kristof	Venter	Goodall	Clinton	Joel	Remnick	Giuliani	Sorkin	Dimon	McCourt
Rashad	X	0.56530	0.59452	0.56268	0.52473	0.605995	0.58188	0.45788	0.60454	0.56416	0.58929	0.61397	0.38059
Newhouse	0.56530	X	0.76004	0.73524	0.70037	0.742985	0.76322	0.46720	0.76373	0.73246	0.76610	0.74892	0.52253
Karr	0.59452	0.76004	X	0.74398	0.69720	0.788840	0.76382	0.55846	0.76832	0.78258	0.79311	0.76429	0.51639
Kristof	0.56268	0.73524	0.74398	X	0.68770	0.783582	0.74718	0.49320	0.75333	0.76001	0.75410	0.74538	0.51926
Venter	0.52473	0.70037	0.69720	0.68770	X	0.737847	0.77987	0.44893	0.77892	0.72764	0.72825	0.72809	0.60091
Goodall	0.60599	0.74298	0.78884	0.78358	0.73784	X	0.81892	0.53168	0.81311	0.82566	0.81092	0.78768	0.57204
Clinton	0.58188	0.76322	0.76382	0.74718	0.77987	0.818923	X	0.50007	0.85201	0.81102	0.79636	0.81300	0.64164
Joel	0.45788	0.46720	0.55846	0.49320	0.44893	0.53165	0.50007	X	0.50336	0.54823	0.59471	0.58806	0.27260
Remnick	0.60454	0.76373	0.76832	0.75333	0.77892	0.81311	0.85201	0.50336	X	0.79466	0.79430	0.80452	0.65072
Giuliani	0.56416	0.73246	0.78258	0.76001	0.72764	0.82566	0.81102	0.54823	0.79466	X	0.81478	0.79870	0.56098
Sorkin	0.58929	0.76610	0.79311	0.75410	0.72825	0.81092	0.79636	0.59471	0.79430	0.81478	X	0.80172	0.52722
Dimon	0.61397	0.74892	0.76429	0.74538	0.72809	0.78768	0.81300	0.58806	0.80452	0.79870	0.80172	X	0.55393
McCourt	0.38059	0.52253	0.51639	0.51926	0.60091	0.57204	0.64164	0.27260	0.65072	0.56098	0.52722	0.55393	X
Avg Similarity	0.55379	0.68901	0.71096	0.69047	0.67837	0.73493	0.73908	0.49703	0.74013	0.72674	0.73091	0.72902	0.52657

Figure 3 - Document Cosine Similarity Matrix

The previously mentioned apparent similarity of our documents was further confirmed by our cosine similarity matrix. With an average inter-document similarity of 67.28% and a inter-document similarity high of 85.2%, we were intrigued by the degree of overlap between documents.

When considering document similarity on an individual basis, we found that Vice President Clinton had the highest average degree of similarity with all other speakers (73.9%) while Billy Joel's was the most unique (49.7%).

Discussion of Commencement Speech Archetype

If one is to examine each of these speeches, these findings are consistent with the content of each. Clinton's speech is rather long at 3162 words and includes references to his early life, the importance of education, and his advice for success.

On the other hand, Joel's is 516 words and downplays the importance of education, him stating "I think this is my fifth or sixth doctorate I'm getting, and I didn't graduate high school" (News Staff). Additionally, Joel is the only speaker to ever break out into song mid-speech, leading commencement attendees in a song to the tune of "Down In New Orleans".

As a result of examining these two speeches from both a statistical and textual approach, it appears as though our findings are consistent. In regards to the archetypal commencement speech, former President Bill Clinton is a promising example of what to expect while Billy Joel demonstrates just how original and unconventional a speech can be.

Limitations & Conclusion

In regard to limitations, we believe that the rather limited number of documents available for our corpus impacted our ability to fully explore the topics present. If we were to have access to the speeches that we were unable to locate, let alone the full collection of Syracuse University speeches in its ~150 years of existence, we would be able to account for much more variance in individual speeches. Additionally, the ability to look at topic trends over time would then present itself.

Over the course of this paper we have detailed how we collected our data, manipulated it, and then put it to use. By utilizing LDA modeling, we were able to examine topics within our corpus in order to understand the common trends within our documents as well as the number of documents associated with each topic. We then designated these trends New Beginnings, Timely Action, and Worldliness based on the tokens present within each.

From there, we examined cosine similarity across our documents and noticed that they were overall rather similar. After examining the average similarity of each paper compared to the rest, we determined that President Bill Clinton's speech contained the most overarching elements of a commencement speech while Billy Joel was the most unique speaker.

Further Research

For further research it would be interesting to compare speeches at Syracuse to speeches given at other large universities. Additionally, with access to speeches from the entire history of Syracuse University, we would enjoy examining the change of topic trends over time and determining the similarity of speeches from past decades to the present.

Works Cited

“LDA - How to Grid Search Best Topic Models?.” *Machine Learning Plus*, 7 May 2018, www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/.

News Staff. “Remarks by Billy Joel at SU/SUNY ESF Commencement.” *SU News*, 15 May 2006, news.syr.edu/blog/2006/05/15/remarks-by-billy-joel-at-susuny-esf-commencement/.

“Past Commencement Speakers.” *Syracuse University Commencement*, commencement.syr.edu/stay-connected/past-commencement-speakers/.

Rehurek, Radim. *gensim Documentation*. Release 0.8.6 ed., 2017.

Scikit-learn developers. *Scikit-Learn User Guide*. Release 0.21.dev0 ed., 2018.

SU News, news.syr.edu/.