Textual Analysis of Daily Princetonian Articles

Yang Song and Prof. Brian Kernighan, Princeton University

Department of Computer Science

Motivation

- 21857 issues of the Daily Princetonian released online in digital format using OCR
- Text from articles contain invaluable data relating to cultural, historical and political changes from a unique Princeton perspective
- Textual analysis allows us to capture and visualize trends quantitively

Prior Approaches

- Sentiment analysis of headlines only
- N-gram comparison with Google Books
- Comparison with Town Topics (local newspaper)
- No comparisons between individual words

Our Project

- Frequency analysis of words using NLTK with the Bag of Words model – plot the relative frequency of words across time in articles
- Word2vec model to examine relationships including semantic similarity between words, word associations and analogies
- Articles extracted from archives as XML – then scraped into text files grouped by year
- Text is tokenized with NLTK and preprocessed – (lowercase, stop-words and suffix removal)
- Gensim (Python implementation of word2vec) used to generate word embeddings

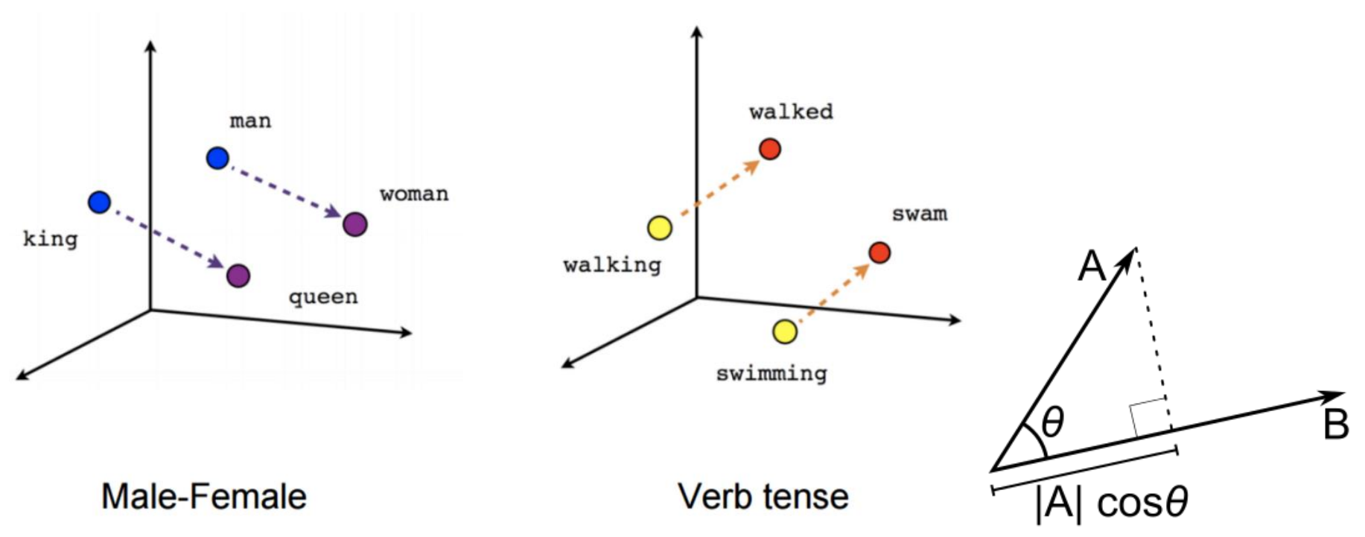
Natural Language Analysis with Python NLTK

Acknowledgments

This work was done for Junior Independent Work for the Computer Science Department in Fall 2018 as part of COS 397.

Word2Vec

- Model used to develop word embeddings (vectors)
- Shallow two-layer neural network
- Transforms each word in text to unique vector (often hundreds of dimensions)
- Two methods – CBOW or skip-grams
- Results often sensitive to parameterization



Male-Female

Verb tense

$$\textbf{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

- Similarity between words measured by the cosine of angle between two vectors
- We can see Colonial is most similar to other sign-in clubs (Charter, Terrace and Quadrangle)
- Analogies between vectors can be determined using addition and subtraction of vectors, e.g. Mathey - Rocky = Whitman - Forbes (sister colleges)
- Vectors can be visualized using dimensionality reduction techniques such as PCA or t-SNE

Word	Cosine Similarity
Yale	0.7954
Cornell	0.7588
Dartmouth	0.7565
Brown	0.7437
Columbia	0.6943

Table 1: Words most similar to Harvard.

Word	Cosine Similarity
Charter	0.6530
Terrace	0.5729
Quadrangle	0.5572
Bicker	0.5538
Barbadian	0.5492

Table 2: Words most similar to Colonial.

Word	Cosine Similarity
Whitman	0.6540
Rockefeller	0.5797
Residential	0.5706
Butler	0.5479

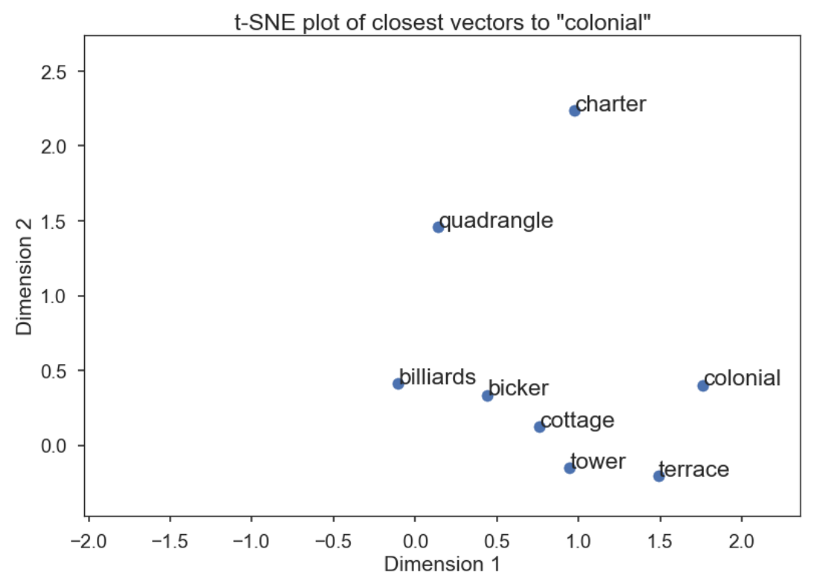
Table 3: Words most similar to Mathey - Rocky + Forbes.

Word	Cosine Similarity
Bulldogs	0.7139
Quakers	0.6901
Bison	0.6731
Colgate	0.6569

Table 4: Words most similar to Crimson - Harvard + Yale.

Word	Cosine Similarity
Quakers	0.6963
Bulldogs	0.6884
Hawks	0.6853
Bears	0.6440

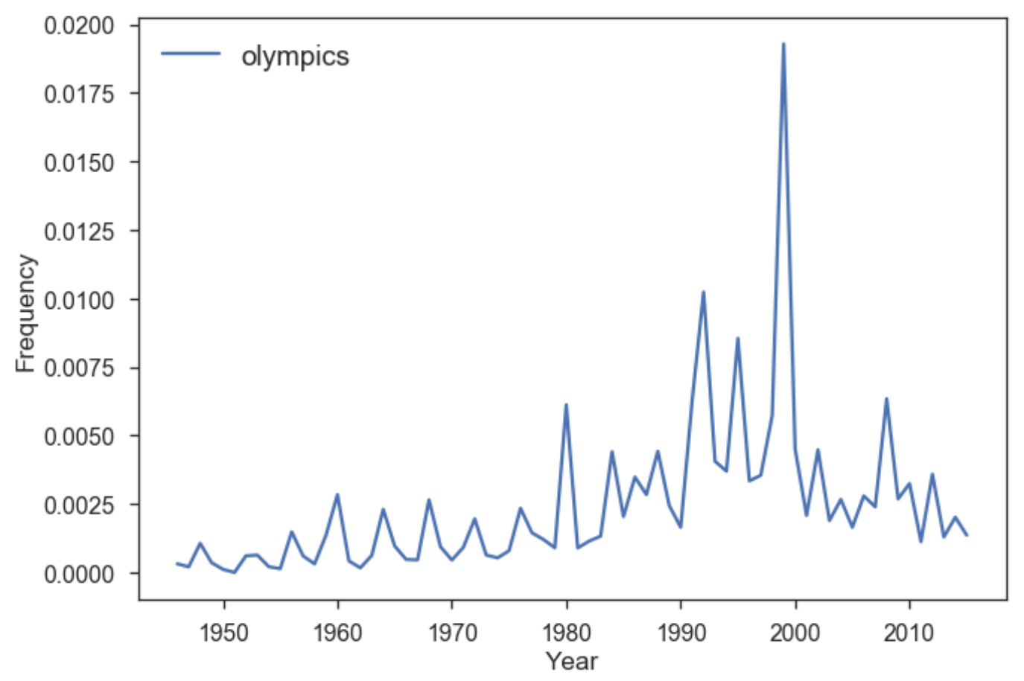
Table 5: Words most similar to Crimson - Harvard + Penn.



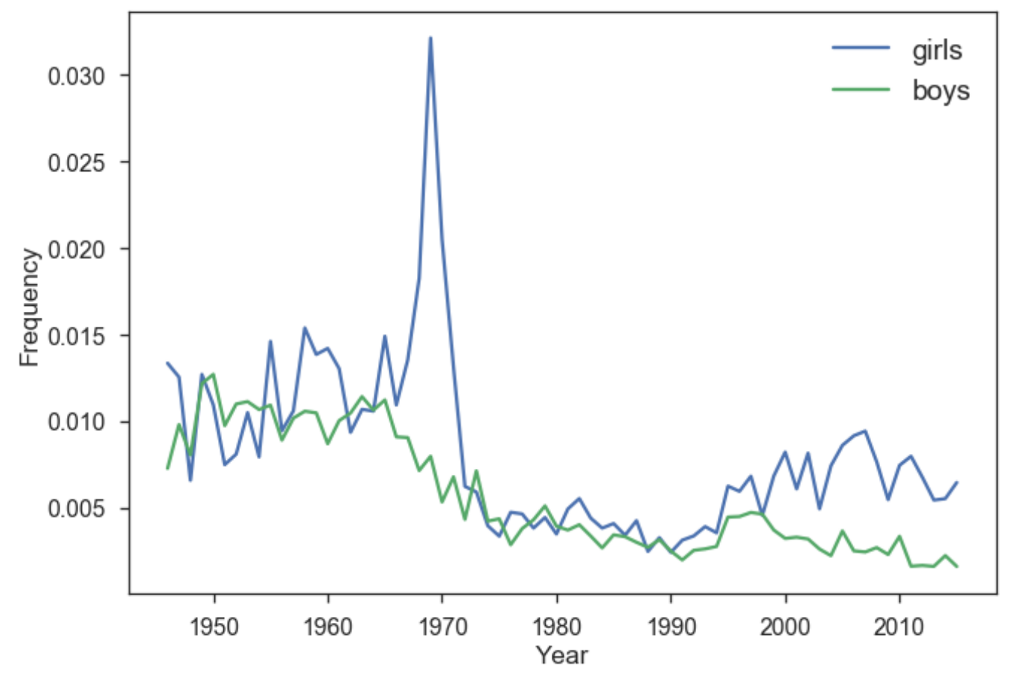
Project Future Work

- Analysis and comparison of headlines and articles using sentence2vec and doc2vec
- Comparison between different campus newspapers
- Web interface for interactive queries by users
- Visualization of change in word vectors over time
- Expand time period to 1800s in results

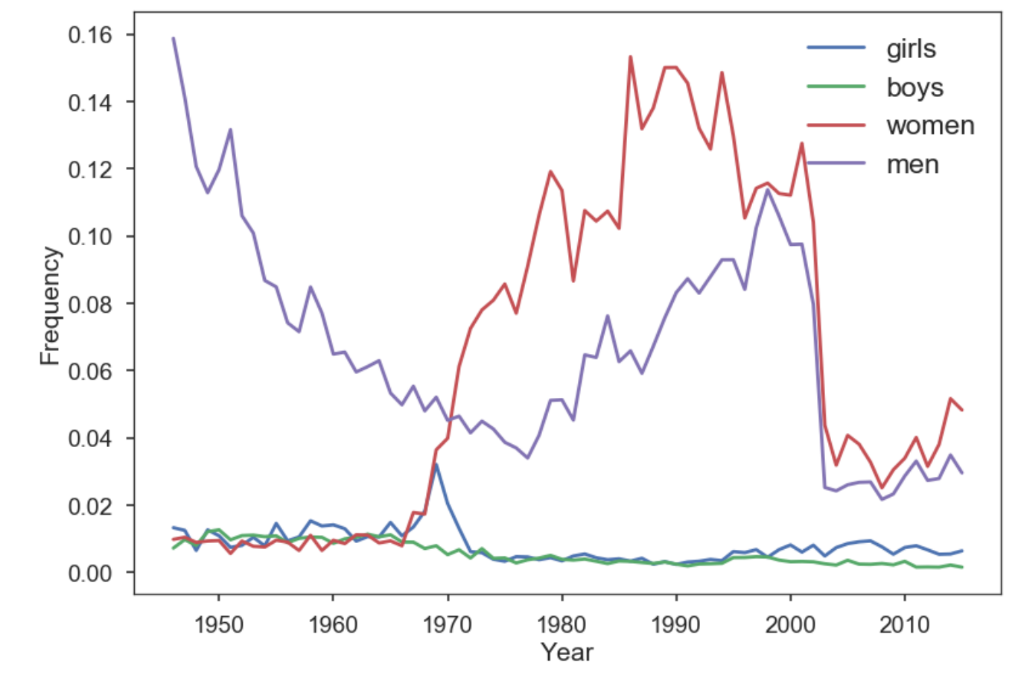
Results



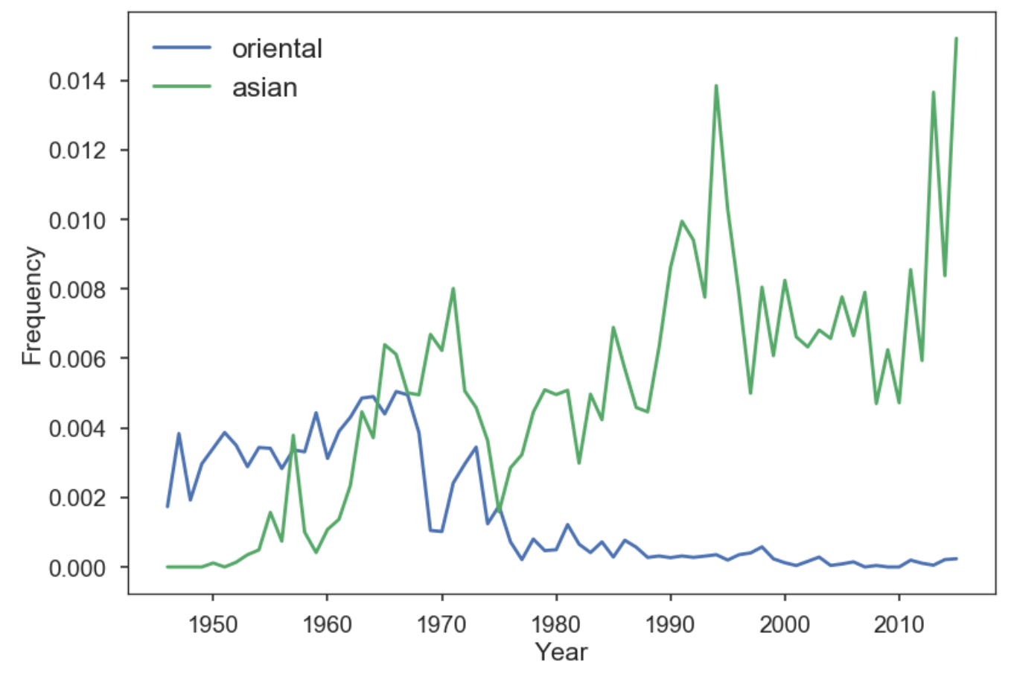
Mentions of ‘Olympics’ peak every 4 years, in 1980 Nude Olympics come into existence until 1999, when they are banned. After 1990s, Winter Olympics alternate 2 years.



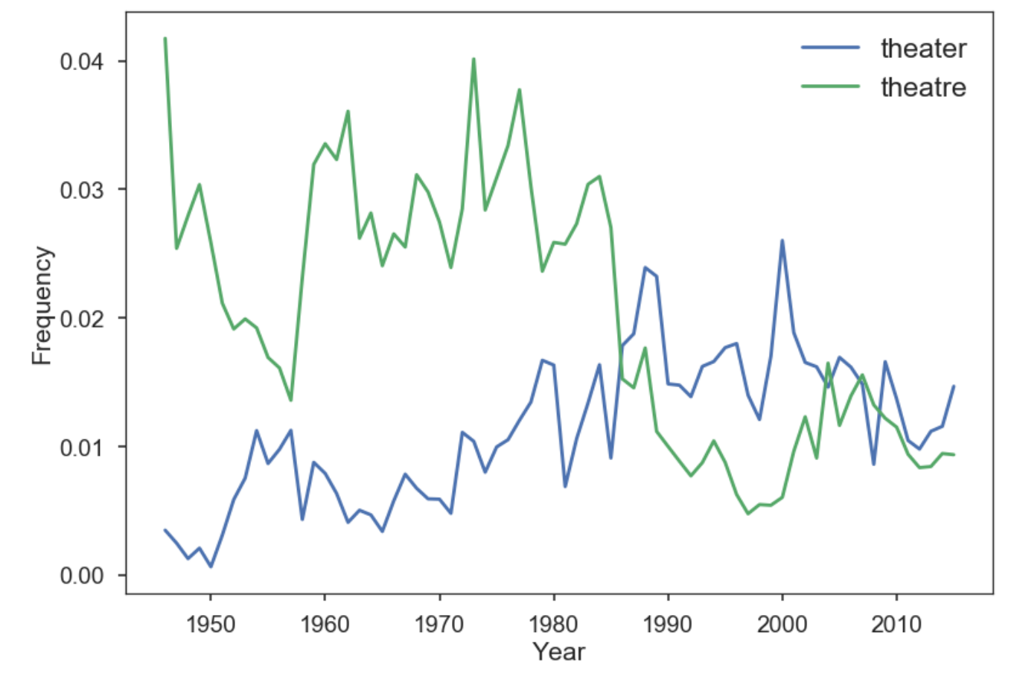
The words ‘girls’ and ‘boys’ have largely been replaced with ‘women’ and ‘men’ over time, but there is a recent resurgence in the use of ‘girls’. The university first admitted women in 1969 (large spike).



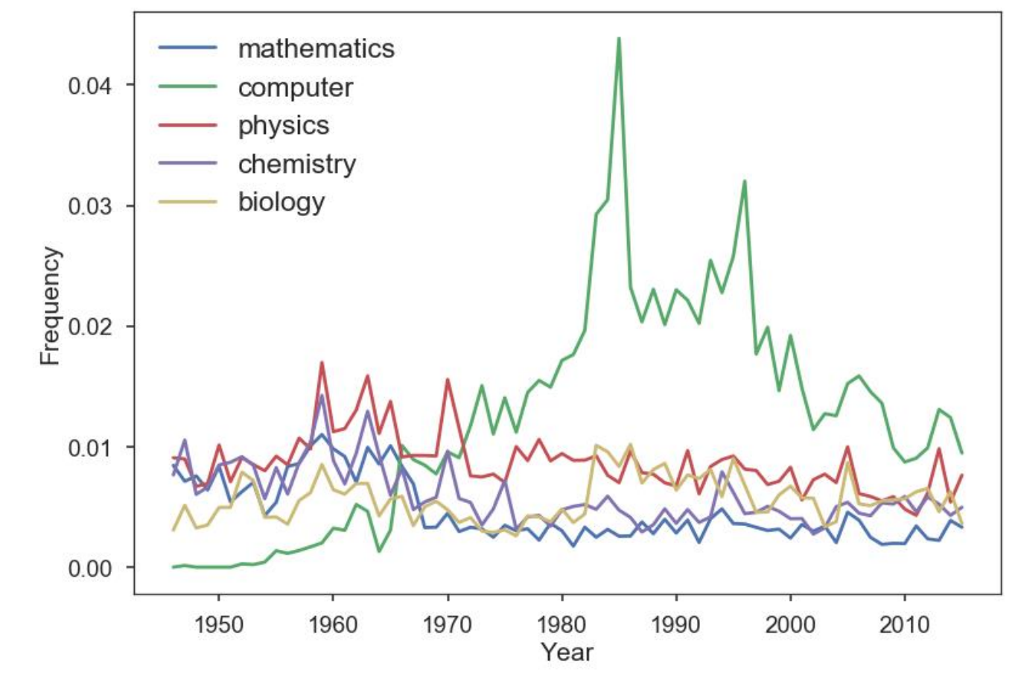
The same graph but now with ‘women’ and ‘men’ added confirms that the use of women/men increased as boys/girls decreased.



The word ‘Oriental’ became replaced with the word ‘Asian’ during the 1970s, and is now virtually non-existent in the Daily Princetonian.



The word ‘theater’ started replacing ‘theatre’ after the 1950s, but now both terms are used interchangeably.



The use of the word ‘computer’ grew rapidly while the other subjects remained relatively steady. (Note the sharp decline in ‘mathematics’ before 1970)