Wikipedia Structure Analysis

Nick Porter, Madhur Pandey, Christian Rachmaninoff
School of Computing
University of Utah

INTRODUCTION

Three Main Questions

- How articles containing similar topics would cluster and their relative distances.
- Exploring how words in these articles were related to each other, and extract some type of semantic meaning
- Which articles were most popular on Wikipedia, and how this related to the links between articles.

Data Preprocessing

Text Processing

- Downloaded the full English Wikipedia dump (58GB)
- Python scripts to parse Wiki-Markup files
- Created a plain text file for each article

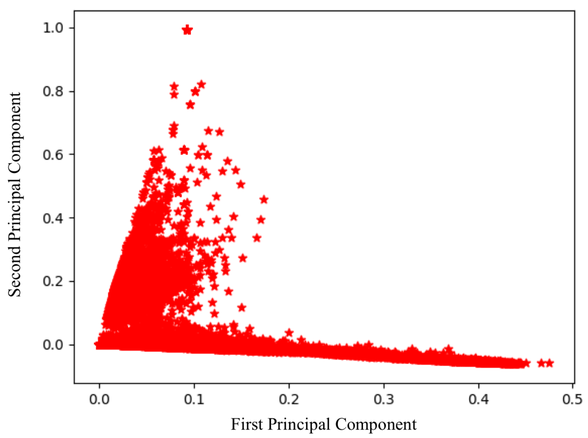
SQL Extraction

- Downloaded the Wikipedia Page Links dump (7GB)
- Created SQL queries to extract page link details

Vectorizing Articles

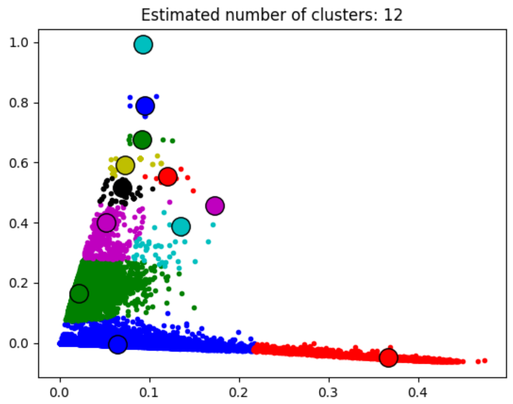
- Removed ‘English’ stop words, converted lowercase
- 3 word grams
- word2vec
- TF-IDF for term weighting
- TruncatedSVD & PCA
- Latent Sementic Analysis (LSA)

120,000 Random Wikipedia Articles Reduced to 2-Dimensions



Clustering Articles

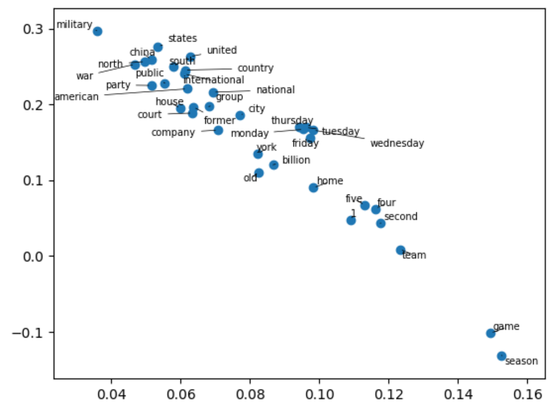
- k-Means
- Mean Shift To Estimate Number of Clusters



Cluster 1	Biological Species
Cluster 2	Education
Cluster 3	Music
Cluster 4	Geographical Information
Cluster 5	Notable People & Dates
Cluster 6	States, Cities, Counties
Cluster 7	Politics
Cluster 8	Films
Cluster 9	Radio & TV Broadcasting
Cluster 10	Sports
Cluster 11	Census Information
Cluster 12	Misc Technology

Embedded Word Nearest Neighbor

- KD-Tree and LSH Comparisons on generated vectors



word	Both	LSH only	KD-Tree only
international	{world regional}	{organization major}	{global european}
war	{conflict occupation}	{end forces}	{invasion wars}
company	{firm companies}	{sold owned}	{subsidiary business}
china	{beijing}	{domestic jiang hu}	{taiwan mainland chinese}
party	{democratic parties}	{socialist liberal}	{coalition opposition}

PageRank

- Implemented Simple PageRank
- Used 1 million edges between articles starting with ‘B’
- Created a tag cloud of the articles with top 100 articles in terms of PageRank.

Conclusions

- KD-Tree and LSH performed equally well for NNS
- Full dataset required to get meaningful results from PageRank

Future Work

- Scale experiments with Spark
- Compare Jaccard distance to link distance
- Compare PageRank to actual page hits
- Inspect outliers in 2D Plot

REFERENCES

- Sci-kit Learn (http://scikit-learn.org/)
- NLTK (http://www.nltk.org/)
- Networkx (http://networkx.readthedocs.io)
- Jurafsky, D. (n.d.). 16. In Speech and Language Processing.