Discovering Princeton’s History:

A textual analysis of collegiate newspaper headlines

David M. Liu ’18
Adviser: Dr. Brian Kernighan

PRINCETON
UNIVERSITY

THE CENTER
FOR DIGITAL
HUMANITIES
@PRINCETON

Motivation

- In 2012, Princeton OCR’ed all 140-years of *Daily Princetonian* articles.
- Scanned images and raw text data were placed onto online archive: http://theprince.princeton.edu
- **Goal:** collect and visualize the historical, cultural, linguistic and political trends embedded in the archive’s text.

Related Work: Quantitative Textual Analysis

- Sentiment Analysis and Classification
  - Computes numerical polarities (positive or negative)
  - Implemented with machine learning or bag-of-words
  - Typically applied to social media text, but also applicable to news and opinion journalism.

- N-gram Distributions
  - Popularized by Google in 2011 with the digitization of 5 million books.
  - Main goal is to explore “culturnomic” trends contained in textual data spanning large periods of time.

Approach

N-gram distributions + Sentiment Analysis

- Plain n-gram distributions do not indicate the source of any trends, must be inferred from the user.

↓

1. N-gram distributions will be generated from headline text in the *Prince* archive.
2. Sample headlines will accompany the visualization. Most positive and negative headlines will be chosen.
3. These headlines will provide context and identify significant articles to further read.

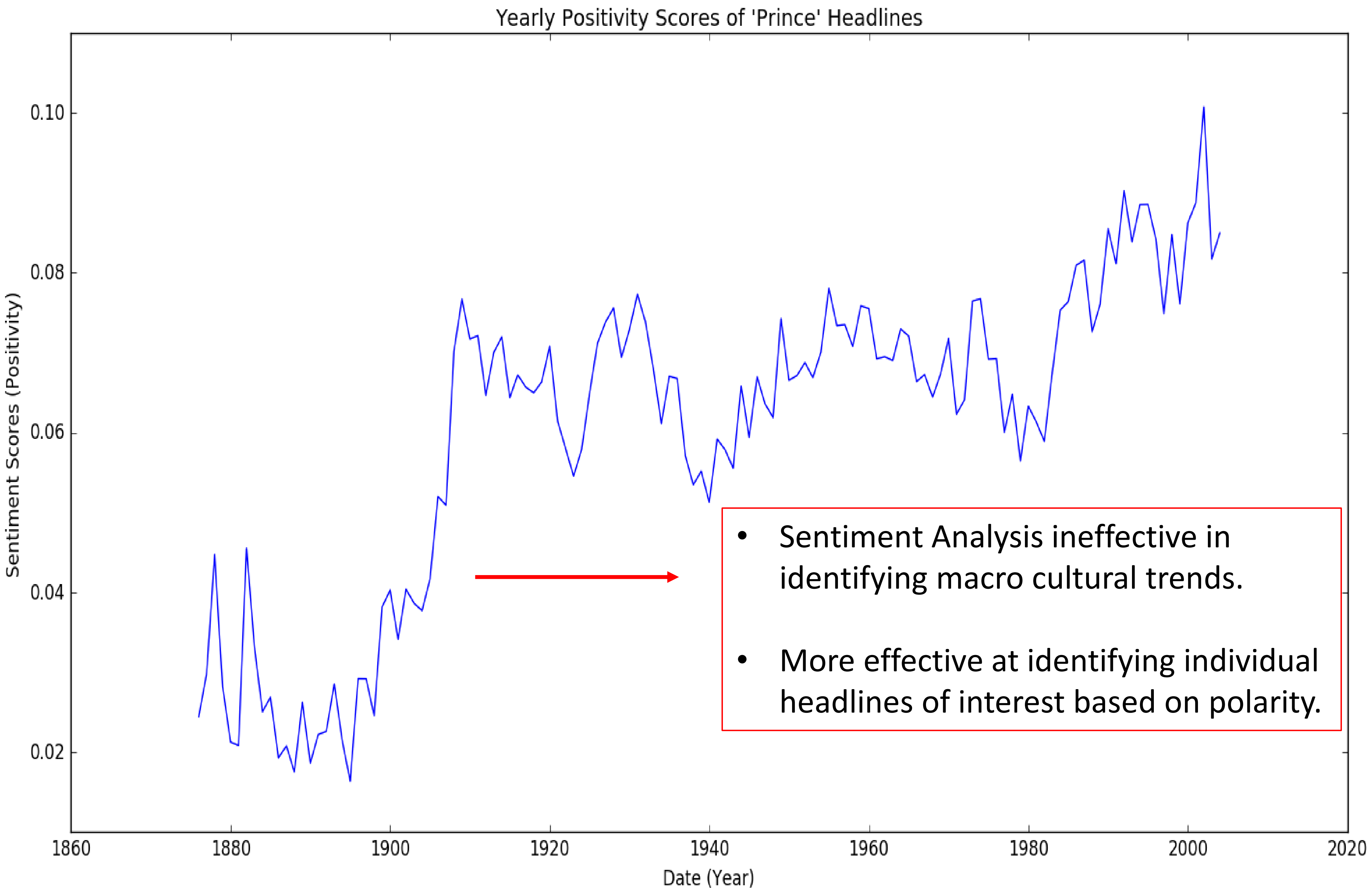
Implementation

1. Scraped all headlines from archive using *BeautifulSoup* and AJAX calls

128	**20K**	**390K**
Years	Issues	Headlines

2. Developed initial visualizations in *matplotlib*, using NLTK's VADER library for sentiment analysis
3. Built final interface into a public webpage using *Flask* to process user queries. Interface available at: http://54.203.14.54:5000/?keywords=men,women

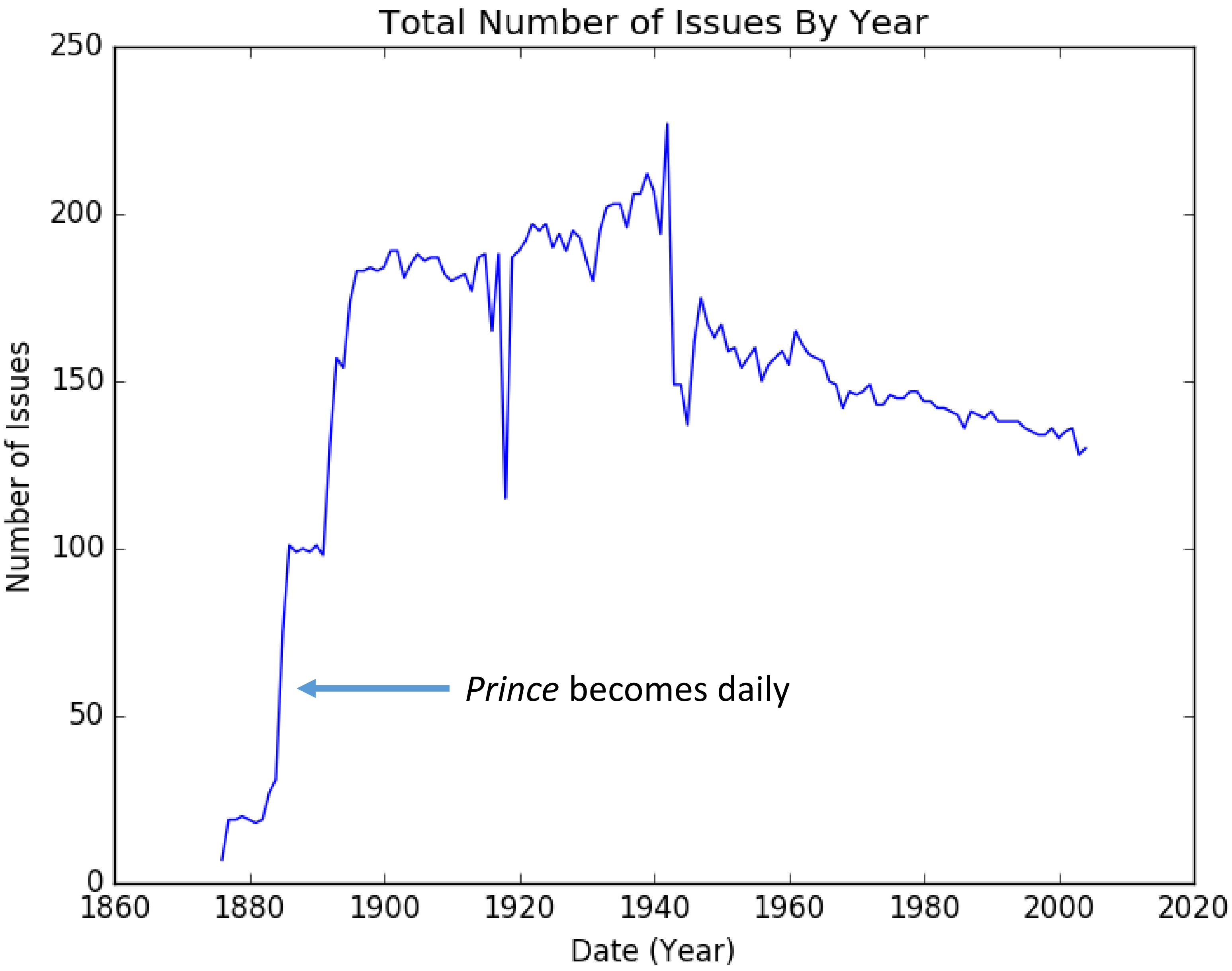
Preliminary: Sentiment Analysis Trends



- Sentiment Analysis ineffective in identifying macro cultural trends.

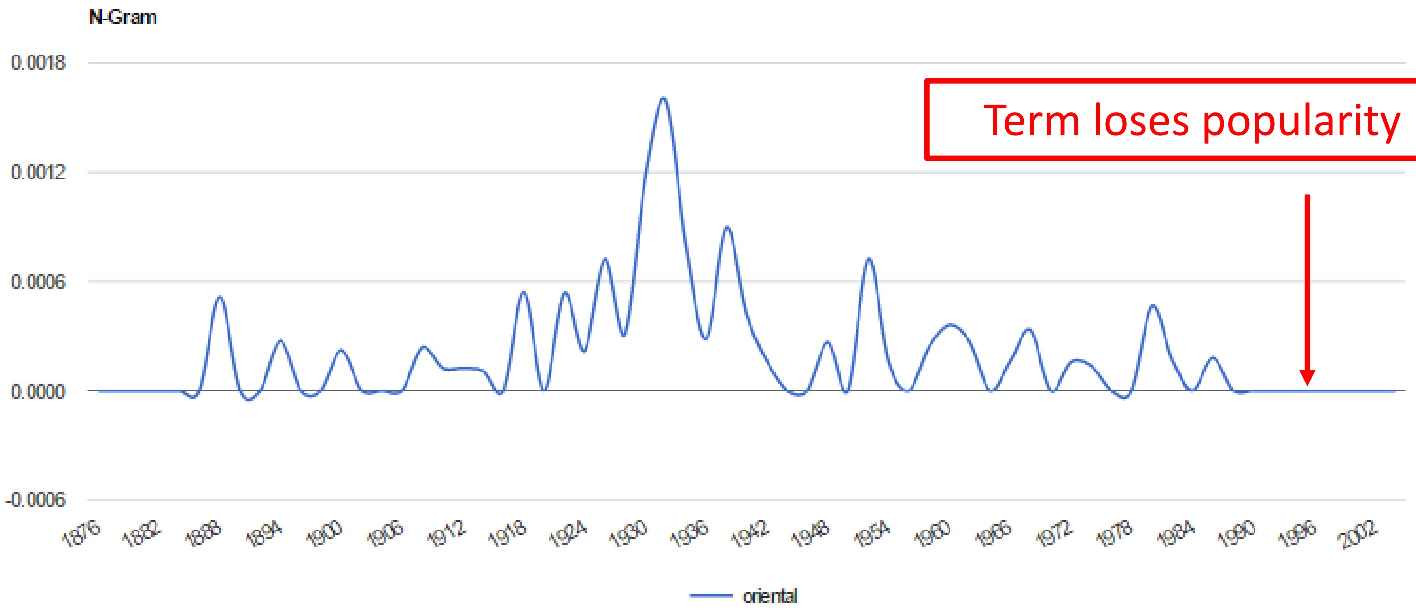
- More effective at identifying individual headlines of interest based on polarity.

Dataset Overview: Headline counts



Prince becomes daily

Results: Linguistic Trends



Term loses popularity by 1990’s

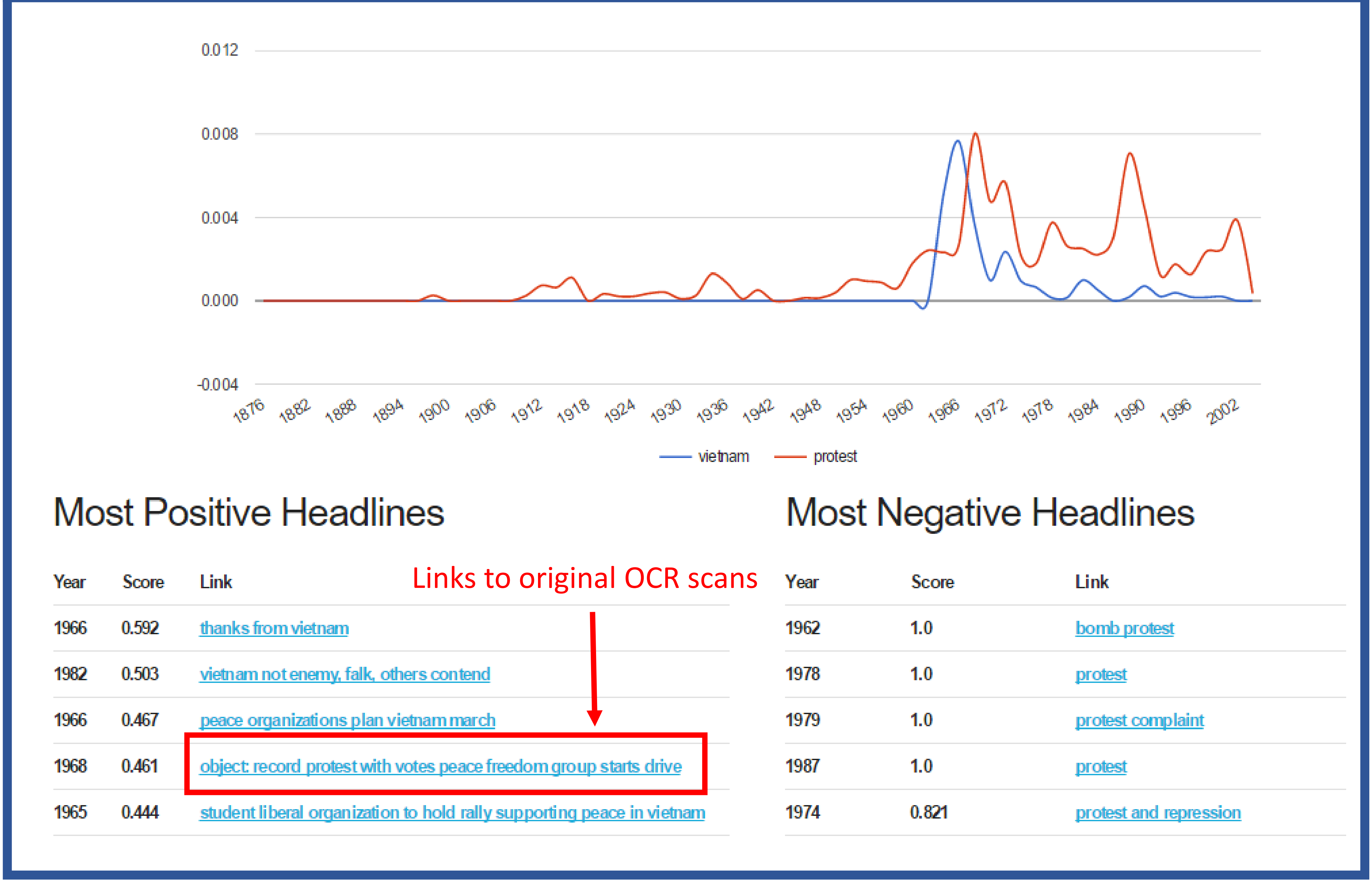
Most Positive Headlines

Year	Score	Link
1986	0.783	oriental romance
1949	0.457	'assurance of peace' essential to iran, states oriental ruler
1926	0.385	oriental comic opera one of triangles best "samarkand," eastern romance of love and intrigue fascinates opening night audiences. 2nd performance tonight novelty of scenic effects contributes to general satisfactory tone of whole performance.
1940	0.3	oriental languages 407 course performs neat disappearing act
1919	0.275	bishop roots discusses problems of far east says america can help oriental relationships by ensuring fair treatment in courts here.

Most Negative Headlines

Year	Score	Link
1968	0.394	student contracts rare oriental flu
1953	0.286	yale social life combines oriental luxury and darkest secrecy
1932	0.278	dr. dennett to explain oriental problems at 8:45
1927	0.266	speaker attributes rising oriental hatred to western fickleness in breaking promises
1975	0.204	upsurge in oriental applications provokes 'encouraging' deviation from national trend

Results: Social and Cultural Trends



Most Positive Headlines

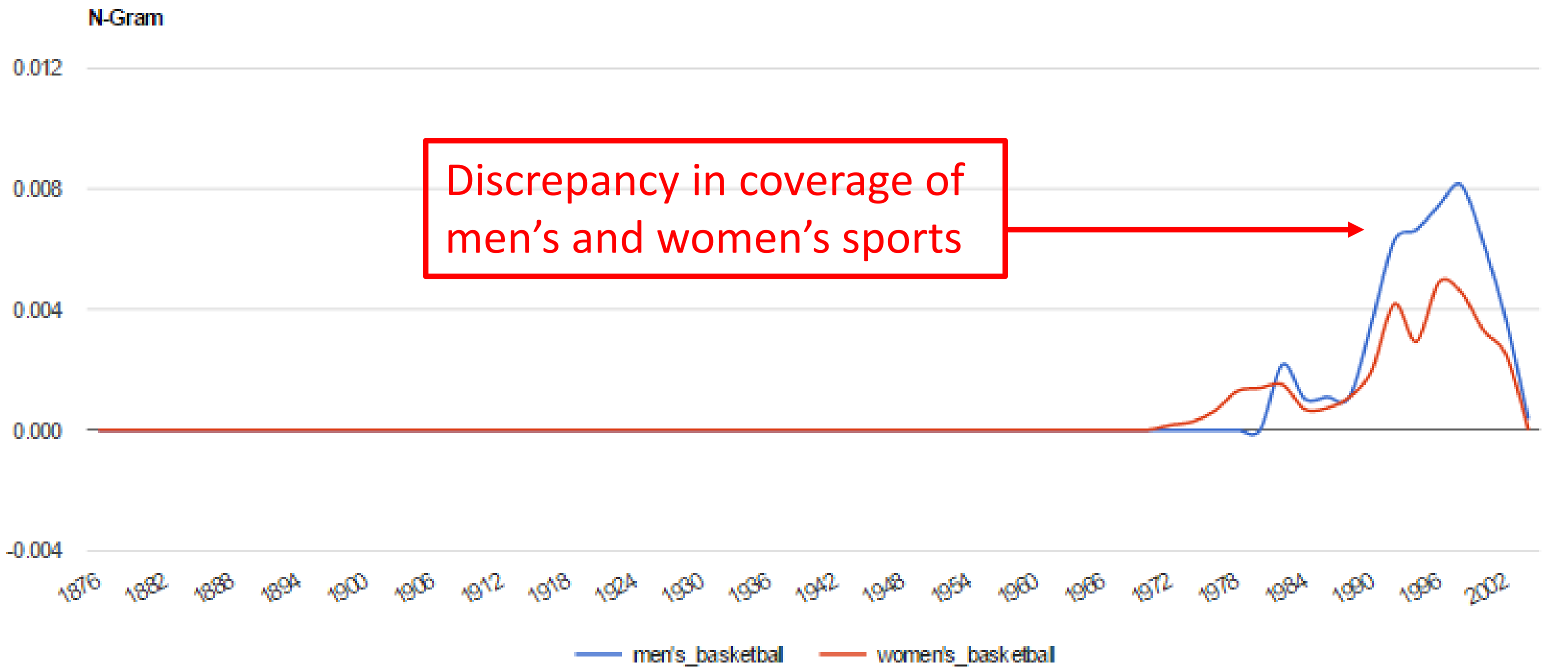
Links to original OCR scans

Year	Score	Link
1966	0.592	thanks from vietnam
1982	0.503	vietnam not enemy, falk, others contend
1966	0.467	peace organizations plan vietnam march
1968	0.461	object: record protest with votes peace freedom group starts drive
1965	0.444	student liberal organization to hold rally supporting peace in vietnam

Most Negative Headlines

Year	Score	Link
1962	1.0	bomb protest
1978	1.0	protest
1979	1.0	protest complaint
1987	1.0	protest
1974	0.821	protest and repression

Results: Journalism Reporting Trends



Discrepancy in coverage of men's and women's sports

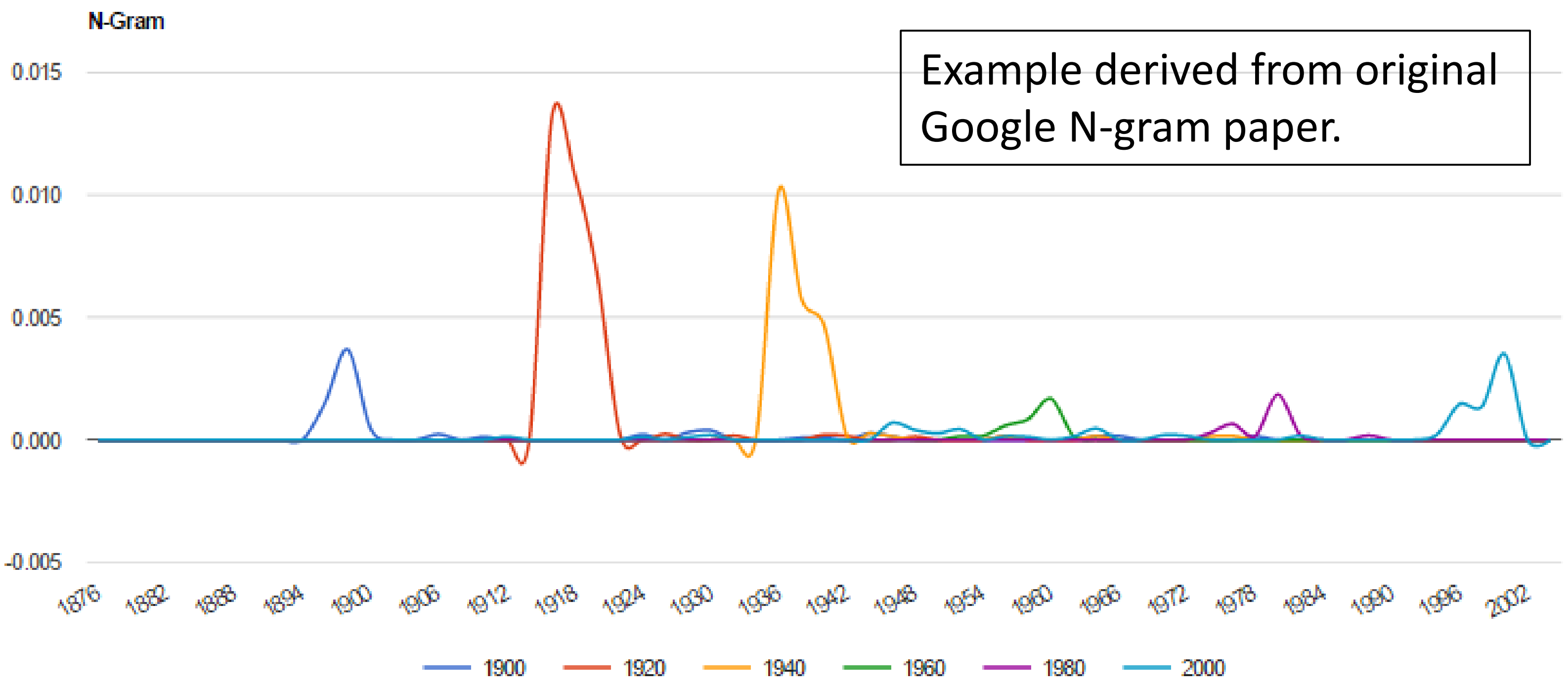
Most Positive Headlines

Year	Score	Link
2002	0.709	men's basketball hopes persia's 80-foot miracle will spark success
2001	0.573	women's basketball gains momentum from exciting first win of season
1997	0.524	all-ivy men's basketball honors
2002	0.508	women's basketball hopes to continue climb to respectability

Most Negative Headlines

Year	Score	Link
1992	0.588	women's basketball struggles in disappointing campaign
1994	0.577	men's basketball falls victim to poor shooting in defeat
1996	0.524	sports shooting problems plague struggling men's basketball
1997	0.504	women's basketball suffers 10th defeat at st. peter's

Results: Perception of Time Trends



Example derived from original Google N-gram paper.

Most Positive Headlines

Year	Score	Link
1900	0.592	1900 football championship.
1937	0.552	purnell wins 1940 post
1976	0.5	welcome class of 1980
1936	0.487	1940 stickmen win by default

Most Negative Headlines

Year	Score	Link
1936	0.467	1940 amateurs fail to appear
2001	0.467	campus crime swells in 2000
1919	0.438	issue 1920 preliminary war records to-day
1937	0.412	rain postpones 1940 golf

Design Choices and Tradeoffs

- Design Choice: scraping only the headlines
  - N-gram perspective: less effective because the corpus is smaller. Query set will be limited.
  - Sentiment perspective: more effective because VADER performs better at the sentence-level.

- Design Choice: casting all text to lowercase
  - Improved n-gram performance by consolidating query terms, increasing the visibility of trends.

Strengths of Sentiment Analysis & N-gram

- Sentiment analysis
  - Successfully classified headlines even though VADER is designed for modern data sets.
  - Fast and efficient computationally
  - Ex: *“colonial gains strength with increased support”* published Sep. 16, 1982. Positivity score = <u>0.839</u>

- N-gram
  - Able to search for both rare and ubiquitous terms
  - N-gram distributions can be compared to each other
  - Less vulnerable to noise in the data
  - User customizable and interactive

Applications and Future Work

Applications:

- Organization and scraping of *Prince* headlines will be useful to the paper's current staff, helping them better understand the publication's history.
- Public n-gram interface can be used by university scholars and public viewers alike

Future Work:

- Expand to even more publications nationwide.
- Develop a metric to rate the quality of an n-gram distribution. Automate discovery of interesting n-gram distributions.
- Create a recommendation system for search queries based on previous searches.

References

[1] Liu, Bing, and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In Mining text data., 415Springer US.

[2] Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. Icwsm 7 (21): 219-22.

[3] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[4] Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. Quantitative Analysis of Culture Using Millions of Digitized Books. Science (Published online ahead of print: 12/16/2010)

[5] Reisman, Dillon. . All the News That's Fit to Change: Insights into a Corpus of 2.5 Million News Headlines, Edited by Joel Reidenberg. Princeton University: Center for Information Technology Policy, 2016.