hci+d lab.

Week 08 • 데이터 저널리즘

Text Analysis Using NLTK & KoNLPy

Joonhwan Lee
human-computer interaction + design lab.

오늘 다룰 내용

- NLTK
- KoNLPy

Text Analysis

Text

- Text - 가장 대표적인 정보의 저장 단위
  - 인류가 만들어내는 텍스트의 양은 현저하게 증가하고 있다
    - WWW
    - Digital Text
    - ...
- Some Facts
  - "The daily New York Times now contains more information that the 17th century man or woman would have encountered in a lifetime." (Wurman, S.A. (1987) Information Anxiety. New York: Doubleday, 32.)
  - "It would take over 200,000 years to read all the Internet, allowing for 30 minutes per document." (Badwin, D. (2001) Information Overload. London: South Bank University.)

Text data란?

- Documents (문서)
  - Articles, books, novels
  - 이메일, 웹페이지, 블로그, 트위터
  - 태그, 코멘트
  - 컴퓨터 프로그램 소스, 로그데이터
- Collection of documents
  - 메시지
  - 소셜네트웍 데이터
  - publications

왜 text를 분석하는가?

- Understanding
  - 문서의 핵심(gist)를 파악하기 위해서
- Grouping
  - 전체를 조망하기 위해 cluster를 만들거나 분류를 하기 위해서
- Compare
  - document collection 을 비교하거나 collection 이 어떻게 시간에 따라 변해왔는지 파악하기 위해서
- Correlate
  - 텍스트에 나타나는 패턴을 다른 데이터와 비교하기 위해서

- Jason Chuang, Stanford University

Text 분석의 사례

- 문서 군집
- 특성 추출
- 문서 요약
- 필터링
- 추천
- 질의응답 시스템

Question

- 어떻게 텍스트와 문서로 부터 정보를 수집하고 사람들이 이해하기 쉽게 텍스트 정보를 보여줄 수 있을까?
  → text mining & information visualization

Tasks & Goals

- Which documents contain text on topic XYZ?
- Which documents are of interest to me?
- Are there other documents that are similar to this one (so they are worthwhile)?
- How are different words used in a document or a document collection?
- What are the main themes and ideas in a document or a collection?
- Which documents have an angry tone?
- How are certain words or themes distributed through a document?
- Identify “hidden” messages or stories in this document collection.
- How does one set of documents differ from another set?
- Quickly gain an understanding of a document or collection in order to subsequently do XYZ.
- Understand the history of changes in a document.
- Find connections between documents.

- John Stasko

Text Visualization

- Text - nominal data
  - ordinal 이나 quantitative data 처럼 그래프로 표현하기 쉽지 않다.
- Then how..?
  - **Frequency** of words
  - **Relationship or structure** of words

Example: Health Care Reform

- Recent history
  - Initiatives by President Clinton
  - Overhaul by President Obama
- Text data
  - News articles
  - Speech transcriptions
  - Legal documents
- What questions might you want to answer?
- What visualizations might help?

- Jason Chuang, Stanford University

Clinton vs. Obama on Health Care

September 9, 2009, 6:59 pm

Bill Clinton on Health Care, 1993

By CATHERINE RAMPELL

Tonight, President Obama will give a much-awaited speech laying out the case for health care reform. Almost exactly 16 years ago, President Bill Clinton did the exact same thing.

Amazingly, in the decade and a half since ambitious plans for "Hillarycare" crumbled, little about the health care reform debate has changed. In fact, President Obama could plagiarize large chunks of Mr. Clinton's health care speech, delivered on Sept. 22, 1993, and few would think the oration sounded dated.

TEXT

Obama's Health Care Speech to Congress

Published: September 9, 2009

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

TWITTER
LINKEDIN
E-MAIL
PRINT
SINGLE PAGE
REPRINTS
SHARE

Prescriptions

A blog from The New York Times that tracks the health care debate as it unfolds.

· More Health Care Overhaul News

Related

In Speech, Obama Will Not Insist on Public Option (September 10, 2009)

Blog

The Caucus

The latest on President Obama, the new administration and other news from Washington and around the nation. Join the discussion.

· More Politics News

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

http://economix.blogs.nytimes.com/2009/09/09/bill-clinton-on-health-care-1993/#more-30335

http://www.nytimes.com/2009/09/10/us/politics/10obama.text.html?_r=0

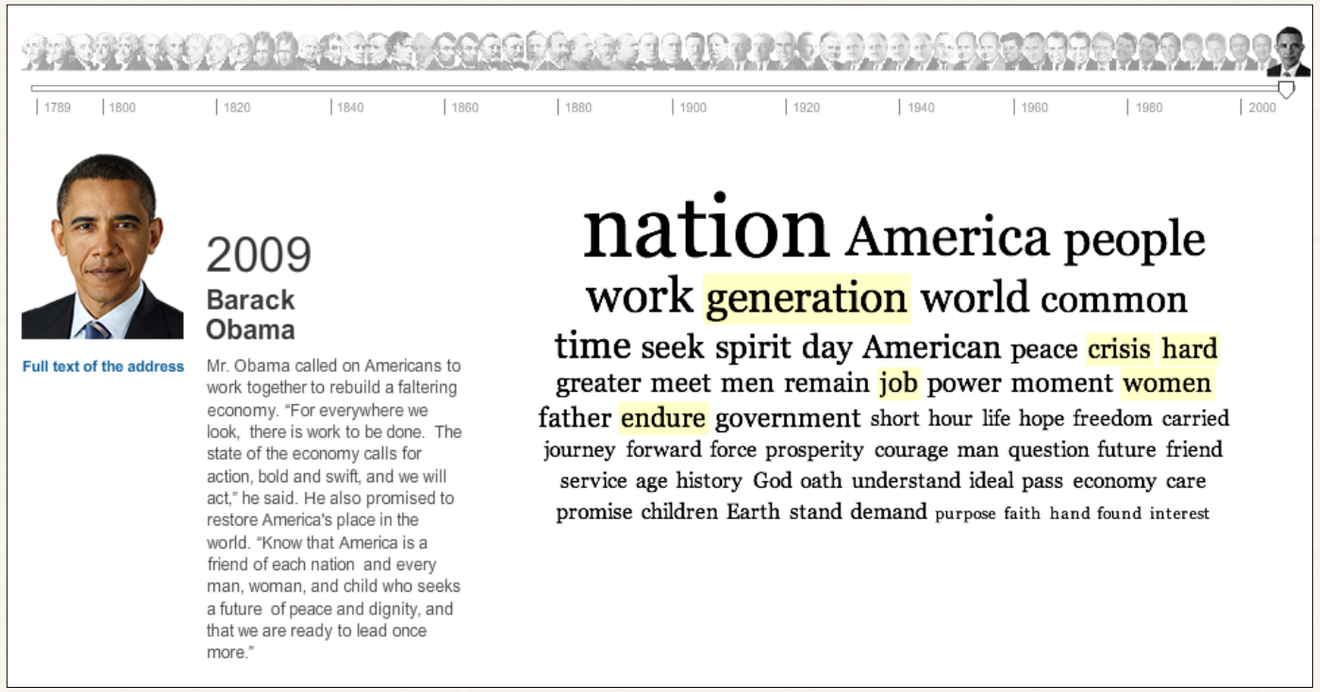
Tag Clouds: Clinton vs. Obama on Health

Mr. Clinton 1993

Mr. Obama 2009

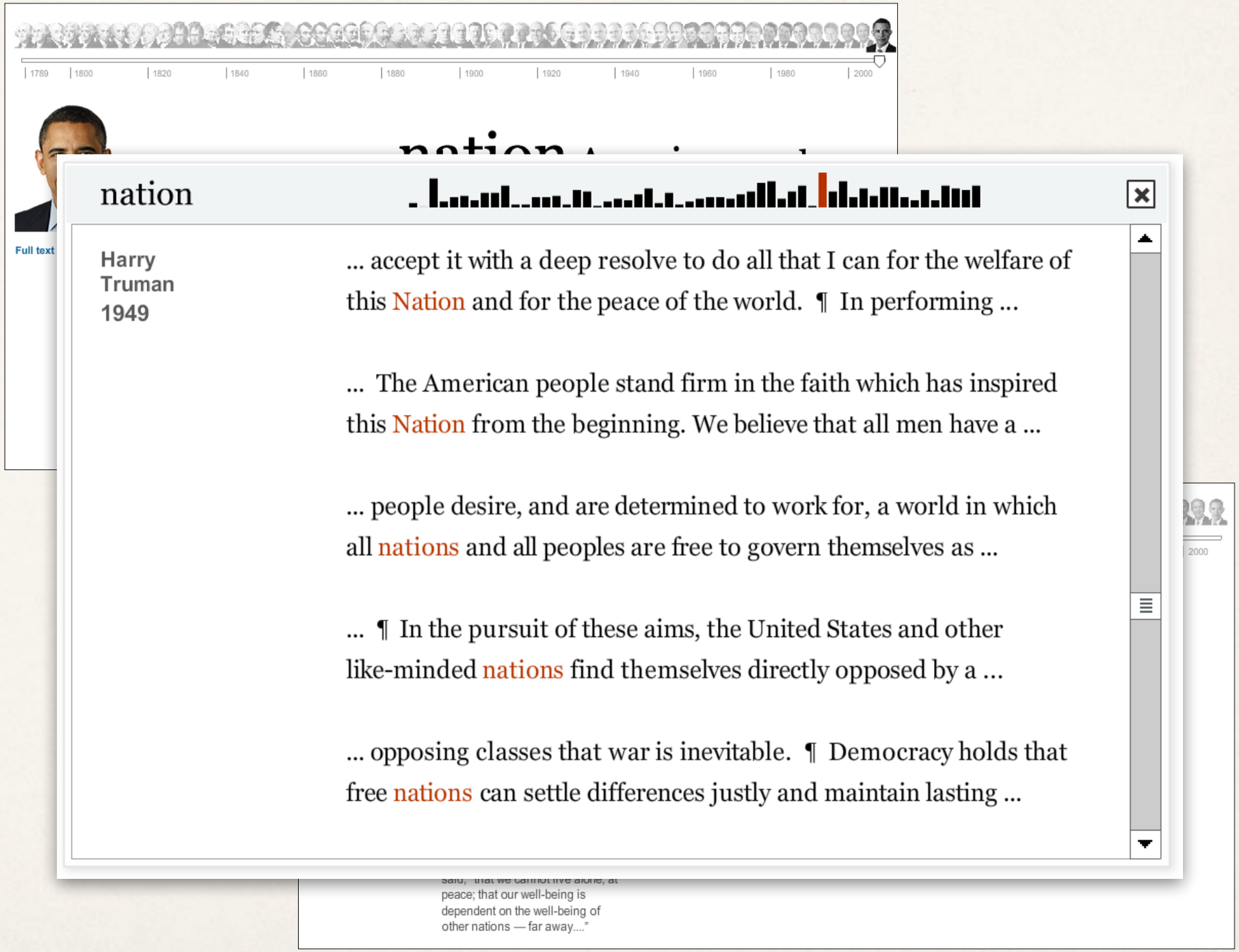
http://economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93/

Inaugural Words: 1789 to the Present



http://www.nytimes.com/interactive/2009/01/17/washington/20090117_ADDRESSES.html?_r=0

Inaugural Words: 1789 to the Present



http://www.nytimes.com/interactive/2009/01/17/washington/20090117_ADDRESSES.html?_r=0

Text As Data

- Word는 nominal data 이지만, 의미나 관계에 따라 non-nominal data로 취급할 수 있다.
  - Correlations: Seoul, Tokyo, Los Angeles, Pittsburgh
  - Order: Sunday, Monday, Tuesday, Wednesday...
  - Hierarchy, antonyms, synonyms, entities and more...
  - 이런 구조를 만들어내기 위해서 → **text processing** 이 필요!
    - 데이터 처리 작업

Text Processing

- Tokenization
  - Segment text into terms.
  - Remove stop words: a, an, the, of, to, be...
  - Numbers and symbols: #superbowl, @monot, what?!!
  - Entities: San Francisco, O’Connor, U.S.A.

- Jason Chuang, Stanford University

Text Processing

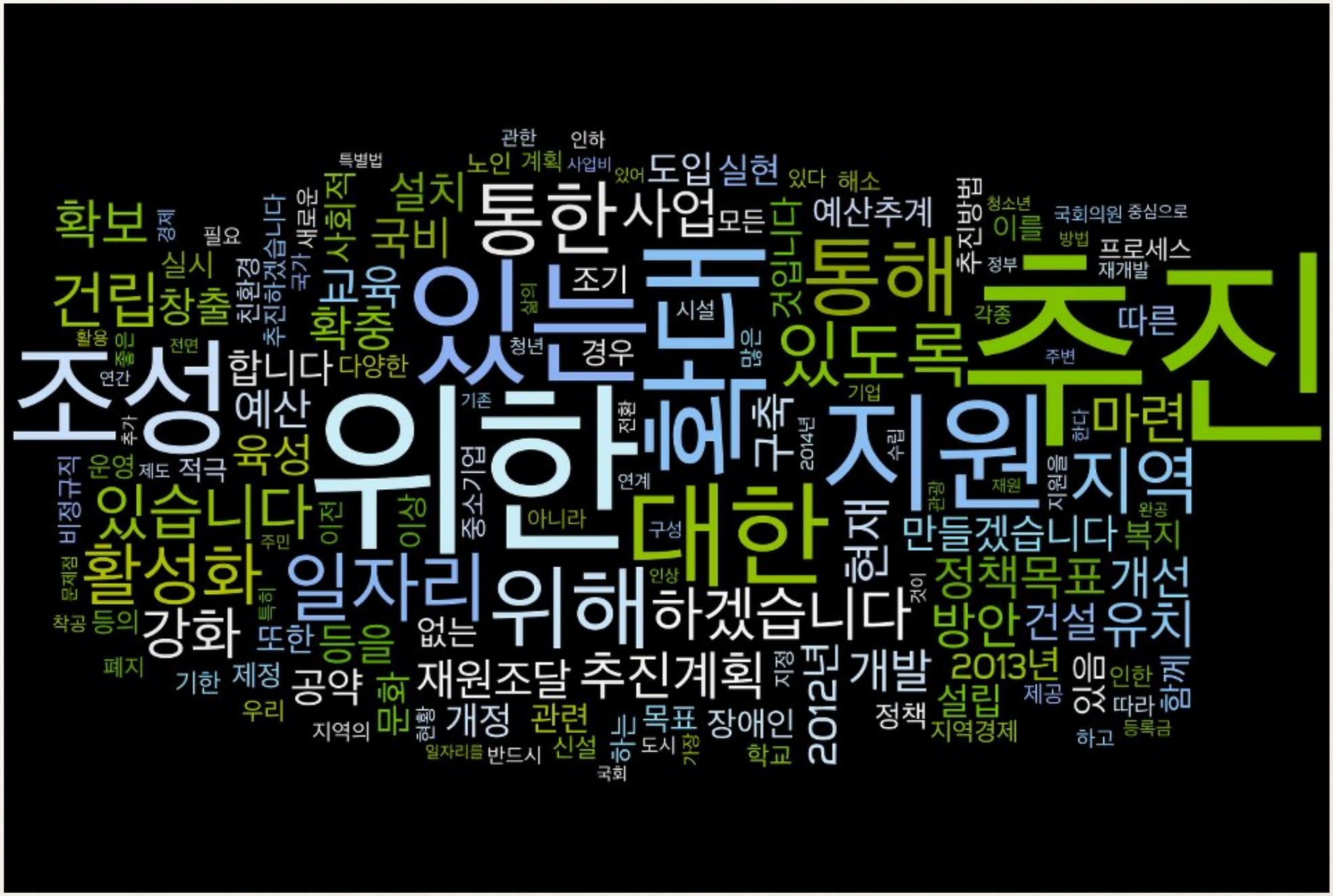
- Stemming
  - Group together different forms of a word
  - Porter stemmer: visualization(s), visualize(s), visually → visual
  - Lemmatization: goes, went, gone → go
- Ordered list of terms
  - appearance order
  - frequency order (sort)

Text Processing

- Stemming → 한국어와 같은 경우, 어미변화가 심해 스테밍 알고리즘으로 처리 곤란
- 형태소 분석기
  - 형태소 분석: 하나의 어절에서 의미를 갖는 최소 단위인 각 형태소를 분석해 내는 것.
  - 문서의 핵심 키워드를 추출하는 기본적인 시스템
    - 예: “영희가 소설책을 읽는다”
      - 자립형태소: 영희, 소설, 책
      - 의존형태소: 가, 을, 읽, 는, 다
      - 실질형태소: 영희, 소설, 책, 읽
      - 형식형태소: 가, 을, 는, 다
  - 초기 텍스트처리는 자립형태소 등을 주로 사용했으나, 최근에는 형태소의 구조적인 관계 및 의미관계까지 고려한 색인어를 추출 → 자연어 검색으로 발전

Text Processing

Wordle: Word Frequency Visualizer



http://www.wordle.net

Questions?