

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
 - ◆ 소셜네트워크 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

Text 분석의 사례

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

Question

- ◆ 어떻게 텍스트와 문서로부터 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.

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?

Clinton vs. Obama on Health Care

September 9, 2009, 6:59 pm

Bill Clinton on Health Care, 1993

By CATHERINE RAMPELL



... a nonpartisan institute that seeks to expand understanding of the presidency, policy, and political history, providing critical insights for the nation's governance challenges.

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.

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.

- TWITTER
- LINKEDIN
- E-MAIL
- PRINT
- SINGLE PAGE
- REPRINTS
- SHARE


<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

Inaugural Words: 1789 to the Present



1789 | 1800 | 1820 | 1840 | 1860 | 1880 | 1900 | 1920 | 1940 | 1960 | 1980 | 2000



2009
Barack Obama

[Full text of the address](#) Mr. Obama called on Americans to work together to rebuild a faltering economy. "For everywhere we look, there is work to be done. The state of the economy calls for action, bold and swift, and we will act," he said. He also promised to restore America's place in the world. "Know that America is a friend of each nation and every man, woman, and child who seeks a future of peace and dignity, and that we are ready to lead once more."

nation America people
work **generation** world common
time seek spirit day American peace **crisis** hard
greater meet men remain **job** power moment **women**
father **endure** government short hour life hope freedom carried
journey forward force prosperity courage man question future friend
service age history God oath understand ideal pass economy care
promise children Earth stand demand purpose faith hand found interest



1789 | 1800 | 1820 | 1840 | 1860 | 1880 | 1900 | 1920 | 1940 | 1960 | 1980 | 2000



1945
Franklin D. Roosevelt

[Full text of the address](#) [Article from the Times archive \(pdf\)](#) In a simple ceremony against a backdrop of World War II, Roosevelt kept his fourth Inaugural Address short at 551 words. Rather than a review of his 12 years in office, it was a plea for an end to the scourge of war and a prayer for lasting peace built on cooperation. "We have learned," he said, "that we cannot live alone, at peace; that our well-being is dependent on the well-being of other nations — far away...."

peace **learn** **men** **test**
world remember perfect people truth courage war
democracy achieve presence God fellow **upward**
understand heart life work live day **strive** friend
gain **mistake** **simple** human purpose service secure principle land
stand century justice great hope American form community believe faith Earth
provide strong office lead fear confidence wish president oath honor member fail
children wisdom importance period constitution instrument word civilization

Inaugural Words: 1789 to the Present

1789 | 1800 | 1820 | 1840 | 1860 | 1880 | 1900 | 1920 | 1940 | 1960 | 1980 | 2000

nation

Harry Truman
1949

... accept it with a deep resolve to do all that I can for the welfare of this **Nation** and for the peace of the world. ¶ In performing ...

... The American people stand firm in the faith which has inspired this **Nation** from the beginning. We believe that all men have a ...

... people desire, and are determined to work for, a world in which all **nations** and all peoples are free to govern themselves as ...

... ¶ In the pursuit of these aims, the United States and other like-minded **nations** find themselves directly opposed by a ...

... opposing classes that war is inevitable. ¶ Democracy holds that free **nations** can settle differences justly and maintain lasting ...

... said, that we cannot live alone, at peace; that our well-being is dependent on the well-being of other nations — far away...."

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.

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 스템밍 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

Korean Morpheme Analyzer Tester

색인어 추출기 분석기

명사만 추출 파일 열기 파일로 저장 분석하기

Contents

존경하는 국민여러분!
700만 해외동포 여러분!

저는 오늘 대한민국의 제18대 대통령에 취임하면서
희망의 새 시대를 열겠다는 각오로 이 자리에 섰습니다.

저에게 이런 막중한 시대적 소명을 맡겨주신
국민 여러분께 깊이 감사드리며,
이 자리에 참석해주신 이명박 대통령과 전직 대통령,
그리고 세계 각국의 경축사절과 내외 귀빈 여러분께도 감사드립니다.

저는 대한민국의 대통령으로서
국민 여러분의 뜻에 부응하여
경제부흥과 국민행복, 문화융성을 이뤄낼 것입니다.

부강하고, 국민 모두가 함께 행복한 대한민국을 만드는데
저의 모든 것을 바치겠습니다.

국민여러분!

오늘의 대한민국은 국민의 노력과
피와 땀으로 이룩된 것입니다.

하면 된다는 국민들의 강한 의지와 저력이
산업화와 민주화를 동시에 이룬

위치	단어	품사	횟수
38	의	JKG	105
85	.	SF	87
105	을	JKO	85
98	ㄴ	ETD	72
128	,	SP	62
3	는	ETD	57
5	국민	NNG	57
161	고	ECE	55
64	를	JKO	49
192	입니다	EFN	49
250	ㄹ	ETD	49
252	이	VCP	49
372	이	JKS	40
82	습니다	EFN	39
48	에	JKM	34
29	는	JX	32
766	있	VV	30
109	어	ECS	29
764	수	NNB	28
62	시대	NNG	21
233	과	JC	21
237	행복	NNG	21
425	우리	NP	21
229	경제	NNG	20
284	만들	VV	19
28	저	NP	18
241	문화	NNG	18
67	것	EPT	17

Console

사전 읽기
완료: 10.196초

단어 추출
전체 단어 수: 2646
완료: 3.372초

316

3.372초

Questions?
