How can we capture multiword expressions?

Seongmin Mun[1], Guillaume Desagulier[2], Anne Lacheret[3] , Kyungwon Lee[4]

[1] Lifemedia Interdisciplinary Program, Ajou University, South Korea
[1,3] UMR 7114 MoDyCo - CNRS, University Paris Nanterre, France
[2] UMR 7114 MoDyCo - University Paris 8, CNRS, University Nanterre
[4] Department of Digital Media, Ajou University, South Korea

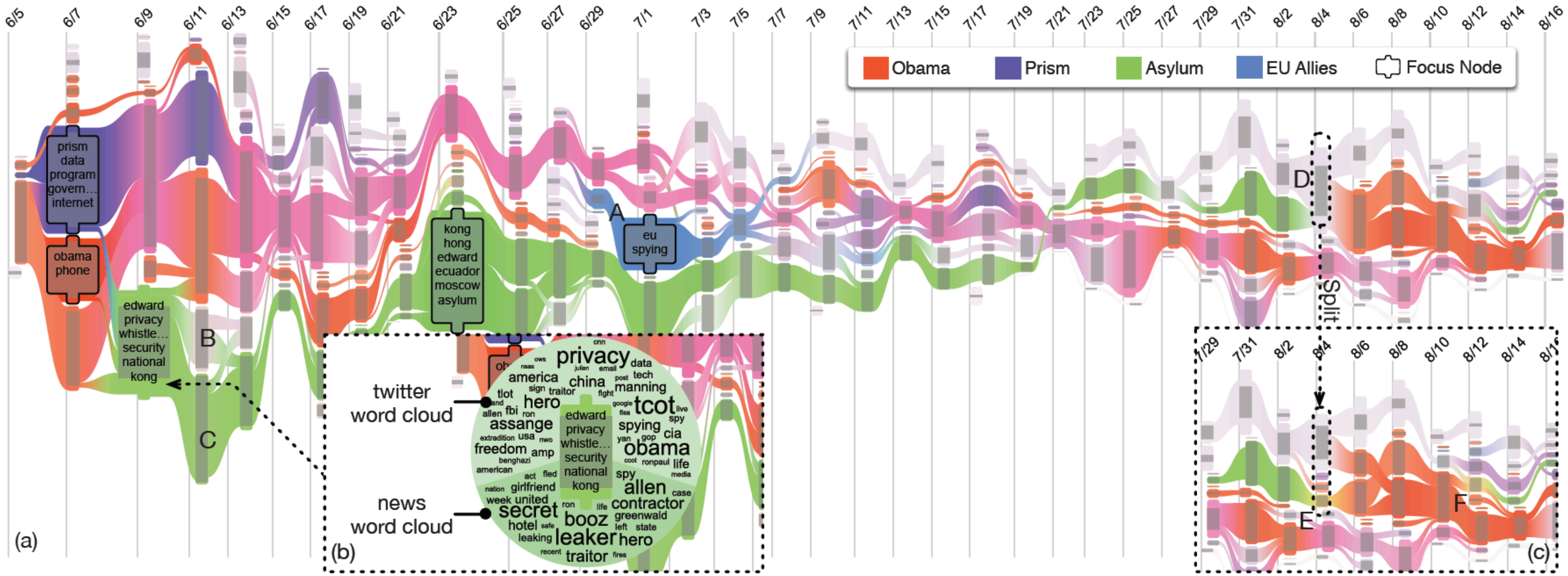
Université Paris Nanterre AJOU UNIVERSITY

Introduction

Topics in a text corpus include features and information.

Analyzing these topics can improve a user's understanding of the corpus.

Previous studies



WEIWEI CUI SHIXIA LIU Z. W. H. W.: How hierarchical topics evolve in large text corpora. In IEEE Tran sactions on Visualization and Computer Graphics (2014), vol. 20, pp. 2281–2290.

Research background and purpose

Topics can be broadly divided into two categories.

Research background and purpose

“With profound gratitude and great humility, I accept your nomination for the presidency of the United States.”

Research background and purpose

“With profound ***gratitude*** and great humility, I accept your nomination for the presidency of the United States.”

Gratitude ➡ meaning that can be expressed in one word

Research background and purpose

“With profound gratitude and great humility, I accept your nomination for the presidency of the ***United States***.”

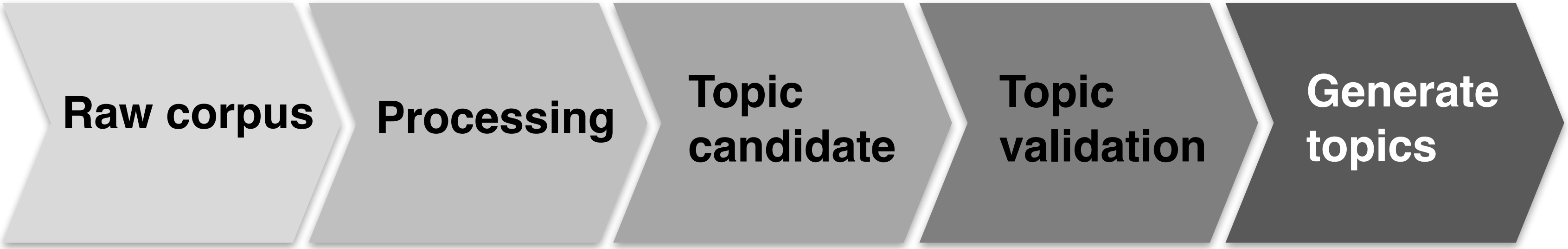
United States → meaning must be described using a combination of words.

Research background and purpose

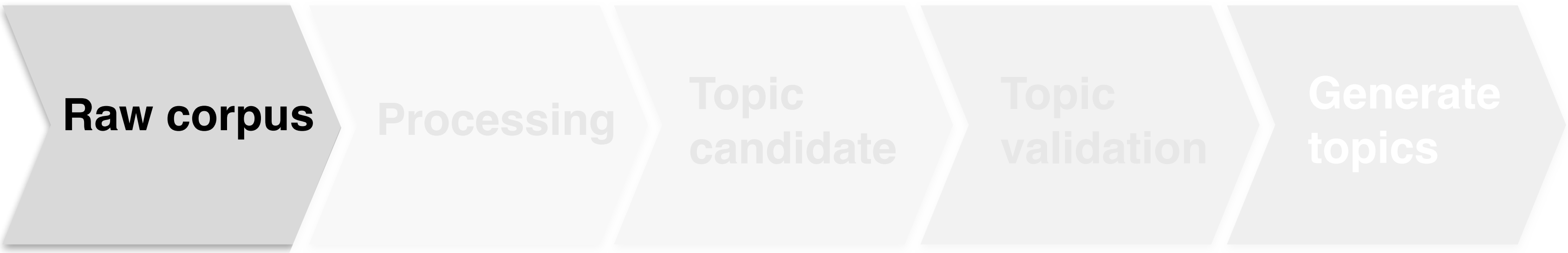
How can we capture multiword expressions?

To this aim, we designed an algorithm.

Data processing



Data processing



Raw corpus

Processing

Topic candidate

Topic validation

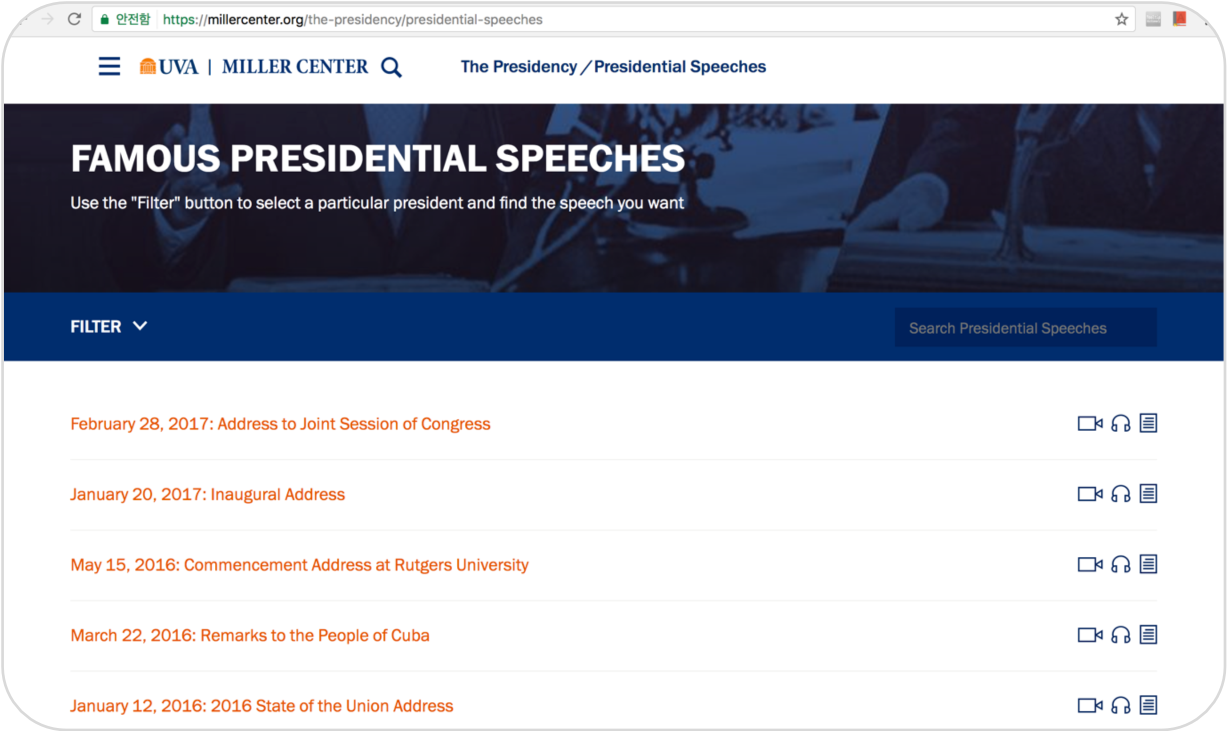
Generate topics

Data processing

Raw corpus
Processing
Topic candidate
Topic validation
Generate topics

Raw corpus

(U.S. president speeches)


안전함 https://millercenter.org/the-presidency/presidential-speeches
UVA | MILLER CENTER
The Presidency / Presidential Speeches
FILTER
Search Presidential Speeches

FAMOUS PRESIDENTIAL SPEECHES

Use the "Filter" button to select a particular president and find the speech you want

February 28, 2017: Address to Joint Session of Congress

January 20, 2017: Inaugural Address

May 15, 2016: Commencement Address at Rutgers University

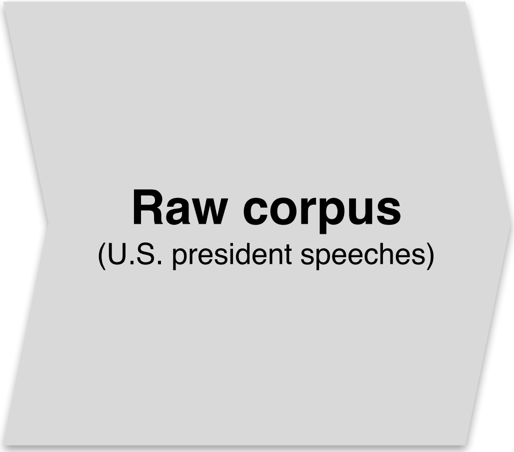
March 22, 2016: Remarks to the People of Cuba

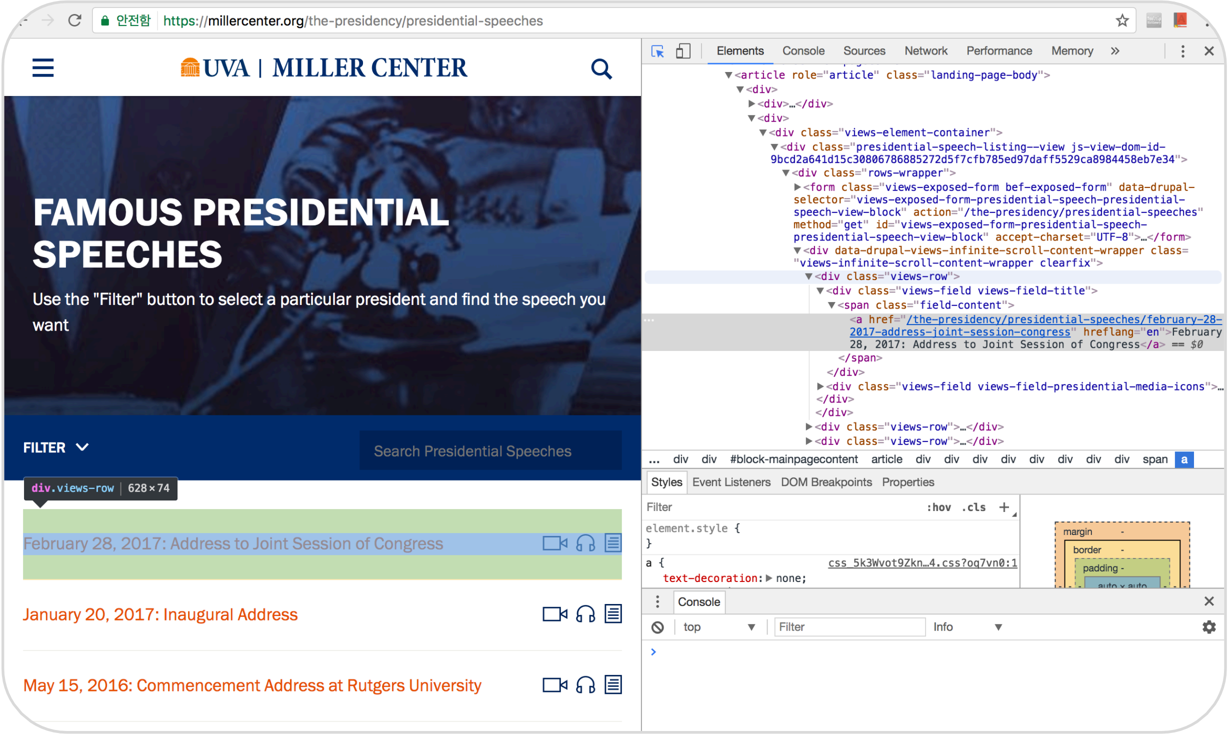
January 12, 2016: 2016 State of the Union Address

https://millercenter.org/the-presidency/presidential-speeches

Data processing

Raw corpus
Processing
Topic candidate
Topic validation
Generate topics


Raw corpus
(U.S. president speeches)



FAMOUS PRESIDENTIAL SPEECHES

Use the "Filter" button to select a particular president and find the speech you want

February 28, 2017: Address to Joint Session of Congress

January 20, 2017: Inaugural Address

May 15, 2016: Commencement Address at Rutgers University

```
▼<article role="article" class="landing-page-body">
  ▼<div>
    ▶<div>…</div>
    ▼<div>
      ▼<div class="views-element-container">
        ▼<div class="presidential-speech-listing--view js-view-dom-id-
        9bcd2a641d15c30806786885272d5f7cfb785ed97daff5529ca8984458eb7e34">
          ▼<div class="rows-wrapper">
            ▶<form class="views-exposed-form bef-exposed-form" data-drupal-
            selector="views-exposed-form-presidential-speech-presidential-
            speech-view-block" action="/the-presidency/presidential-speeches"
            method="get" id="views-exposed-form-presidential-speech-
            presidential-speech-view-block" accept-charset="UTF-8">…</form>
            ▼<div data-drupal-views-infinite-scroll-content-wrapper class=
            "views-infinite-scroll-content-wrapper clearfix">
              ▼<div class="views-row">
                ▼<div class="views-field views-field-title">
                  ▼<span class="field-content">
                      <a href="/the-presidency/presidential-speeches/february-28-
                      2017-address-joint-session-congress" hreflang="en">February
                      28, 2017: Address to Joint Session of Congress</a> == $0
                    </span>
                  </div>
                ▶<div class="views-field views-field-presidential-media-icons">…
                </div>
                </div>
              ▶<div class="views-row">…</div>
              ▶<div class="views-row">…</div>
```

… div div #block-mainpagecontent article div div div div div div div div span a

```
element.style {
}
a {                              css_5k3Wvot9Zkn…4.css?og7vn0:1
  text-decoration: ▶ none;
}
```

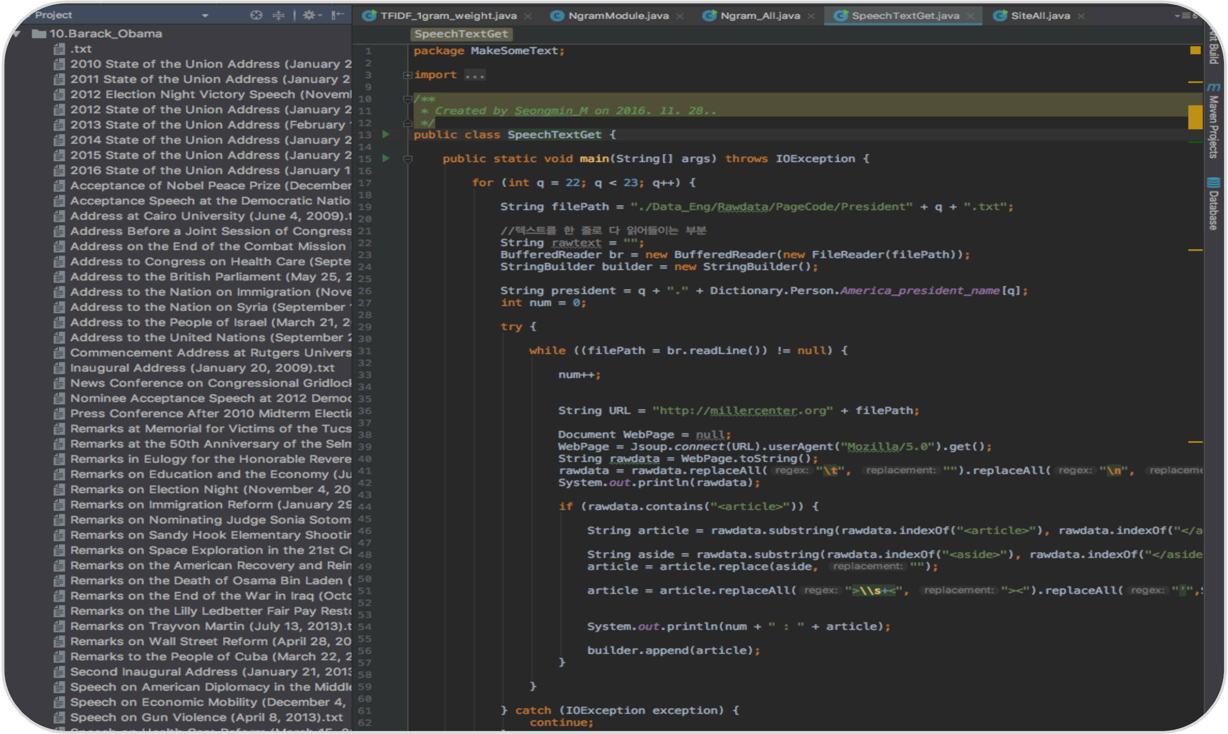
Université Paris Nanterre

AJOU UNIVERSITY

Data processing

Raw corpus

(U.S. president speeches)



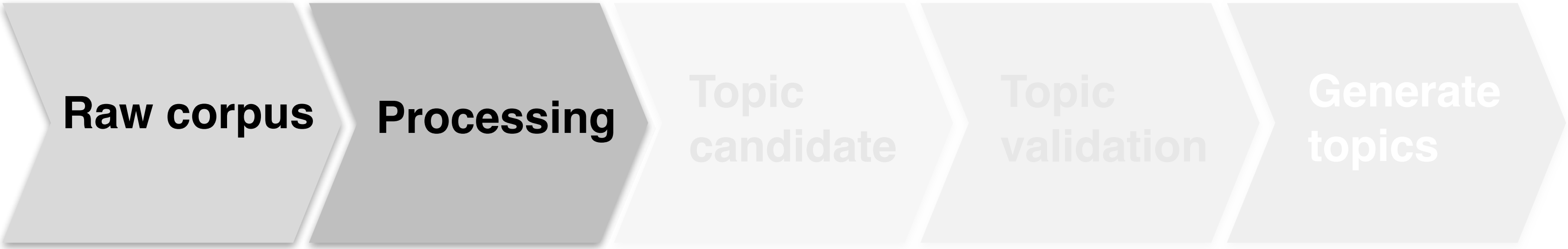
```
10.Barack_Obama
    .txt
    2010 State of the Union Address (January 2
    2011 State of the Union Address (January 2
    2012 Election Night Victory Speech (Novemb
    2012 State of the Union Address (January 2
    2013 State of the Union Address (February
    2014 State of the Union Address (January 2
    2015 State of the Union Address (January 2
    2016 State of the Union Address (January 1
    Acceptance of Nobel Peace Prize (December
    Acceptance Speech at the Democratic Natio
    Address at Cairo University (June 4, 2009).t
    Address Before a Joint Session of Congress
    Address on the End of the Combat Mission
    Address to Congress on Health Care (Septe
    Address to the British Parliament (May 25, 2
    Address to the Nation on Immigration (Nove
    Address to the Nation on Syria (September
    Address to the People of Israel (March 21, 2
    Address to the United Nations (September 2
    Commencement Address at Rutgers Univers
    Inaugural Address (January 20, 2009).txt
    News Conference on Congressional Gridlock
    Nominee Acceptance Speech at 2012 Democ
    Press Conference After 2010 Midterm Electi
    Remarks at Memorial for Victims of the Tucs
    Remarks at the 50th Anniversary of the Selm
    Remarks in Eulogy for the Honorable Revere
    Remarks on Education and the Economy (Ju
    Remarks on Election Night (November 4, 20
    Remarks on Immigration Reform (January 29
    Remarks on Nominating Judge Sonia Sotom
    Remarks on Sandy Hook Elementary Shootin
    Remarks on Space Exploration in the 21st Ce
    Remarks on the American Recovery and Rein
    Remarks on the Death of Osama Bin Laden (
    Remarks on the End of the War in Iraq (Octo
    Remarks on the Lilly Ledbetter Fair Pay Resto
    Remarks on Trayvon Martin (July 13, 2013).t
    Remarks on Wall Street Reform (April 28, 20
    Remarks to the People of Cuba (March 22, 2
    Second Inaugural Address (January 21, 2013
    Speech on American Diplomacy in the Middle
    Speech on Economic Mobility (December 4,
    Speech on Gun Violence (April 8, 2013).txt
```

```java
package MakeSomeText;

import ...

/**
 * Created by Seongmin_M on 2016. 11. 28..
 */
public class SpeechTextGet {

    public static void main(String[] args) throws IOException {

        for (int q = 22; q < 23; q++) {

            String filePath = "./Data_Eng/Rawdata/PageCode/President" + q + ".txt";

            //텍스트를 한 줄로 다 읽어들이는 부분
            String rawtext = "";
            BufferedReader br = new BufferedReader(new FileReader(filePath));
            StringBuilder builder = new StringBuilder();

            String president = q + "." + Dictionary.Person.America_president_name[q];
            int num = 0;

            try {

                while ((filePath = br.readLine()) != null) {

                    num++;


                    String URL = "http://millercenter.org" + filePath;

                    Document WebPage = null;
                    WebPage = Jsoup.connect(URL).userAgent("Mozilla/5.0").get();
                    String rawdata = WebPage.toString();
                    rawdata = rawdata.replaceAll( regex: "\t",  replacement: "").replaceAll( regex: "\n",  replaceme
                    System.out.println(rawdata);

                    if (rawdata.contains("<article>")) {

                        String article = rawdata.substring(rawdata.indexOf("<article>"), rawdata.indexOf("</a

                        String aside = rawdata.substring(rawdata.indexOf("<aside>"), rawdata.indexOf("</aside
                        article = article.replace(aside,  replacement: "");

                        article = article.replaceAll( regex: ">\\s+<",  replacement: "><").replaceAll( regex: "'",


                        System.out.println(num + " : " + article);

                        builder.append(article);
                    }

                }

            } catch (IOException exception) {
                continue;
```

Université
Paris Nanterre

AJOU UNIVERSITY

Data processing



Data processing

Raw corpus → Processing → Topic candidate → Topic validation → Generate topics

Processing

- N-grams
- POS tagging

Pre-processing

- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

N-gram method is a contiguous sequence of ***N*** items from a given sequence of text.

Data processing

Raw corpus → **Processing** → Topic candidate → Topic validation → Generate topics

Processing

- N-grams
- POS tagging

Pre-processing

- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

“Time flies like an arrow.”

Data processing

Raw corpus

Processing

Topic candidate

Topic validation

Generate topics

Processing

- N-grams
- POS tagging

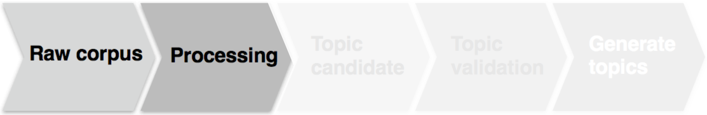
Pre-processing

- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

"Time flies like an arrow."

Unigram : Time, flies, like, an, arrow.
Bigram : Time flies, flies like, like an, an arrow.
Trigram : Time flies like, flies like an, like an arrow.

Data processing

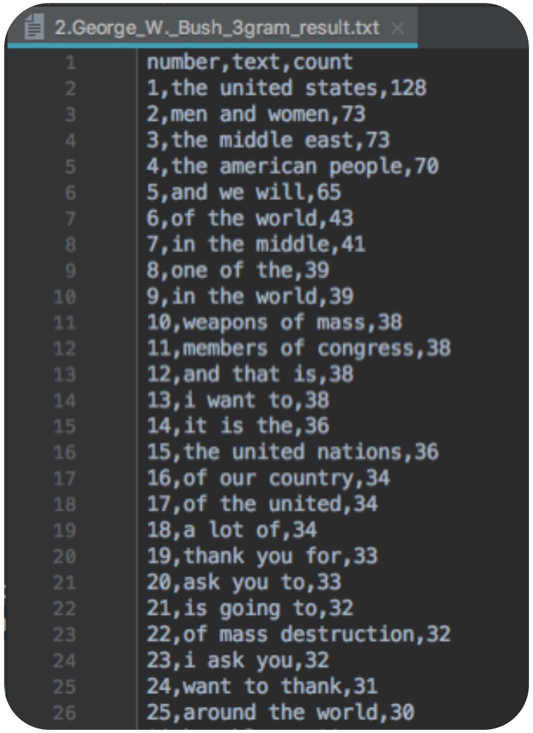


Processing

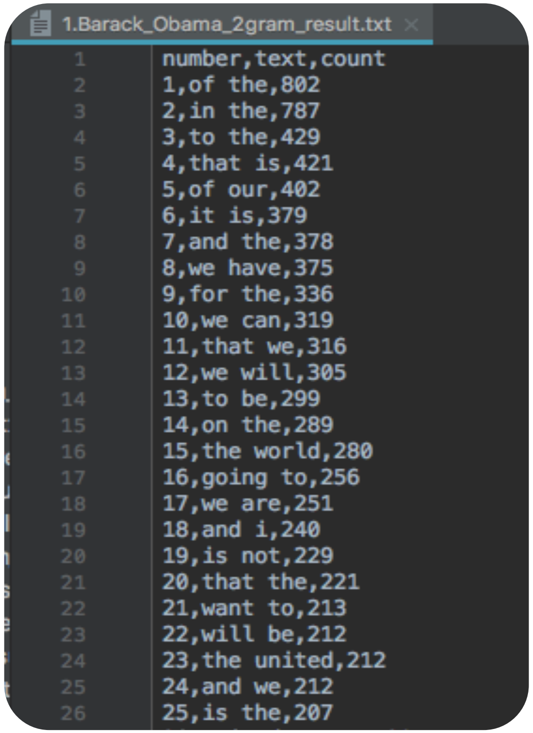
- N-grams
- POS tagging

Pre-processing

- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

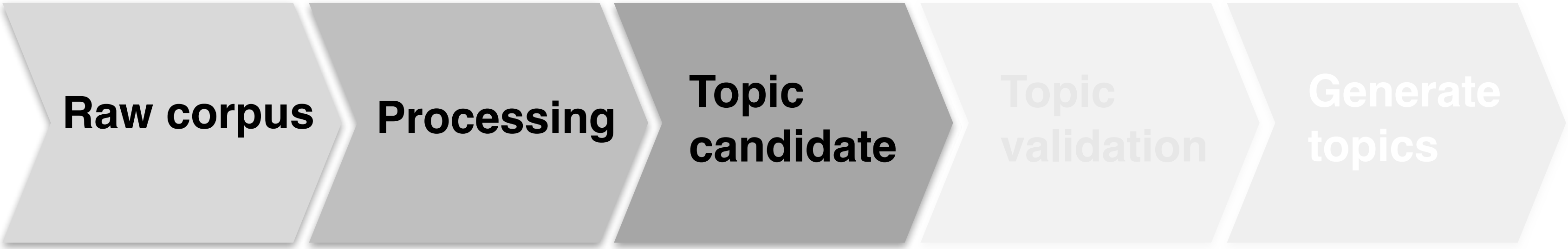


```
number,text,count
1,the united states,128
2,men and women,73
3,the middle east,73
4,the american people,70
5,and we will,65
6,of the world,43
7,in the middle,41
8,one of the,39
9,in the world,39
10,weapons of mass,38
11,members of congress,38
12,and that is,38
13,i want to,38
14,it is the,36
15,the united nations,36
16,of our country,34
17,of the united,34
18,a lot of,34
19,thank you for,33
20,ask you to,33
21,is going to,32
22,of mass destruction,32
23,i ask you,32
24,want to thank,31
25,around the world,30
```

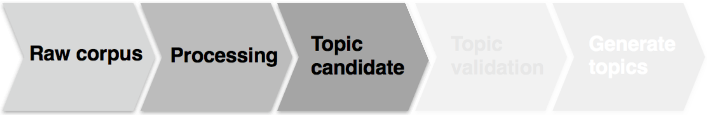


```
number,text,count
1,of the,802
2,in the,787
3,to the,429
4,that is,421
5,of our,402
6,it is,379
7,and the,378
8,we have,375
9,for the,336
10,we can,319
11,that we,316
12,we will,305
13,to be,299
14,on the,289
15,the world,280
16,going to,256
17,we are,251
18,and i,240
19,is not,229
20,that the,221
21,want to,213
22,will be,212
23,the united,212
24,and we,212
25,is the,207
```

Data processing

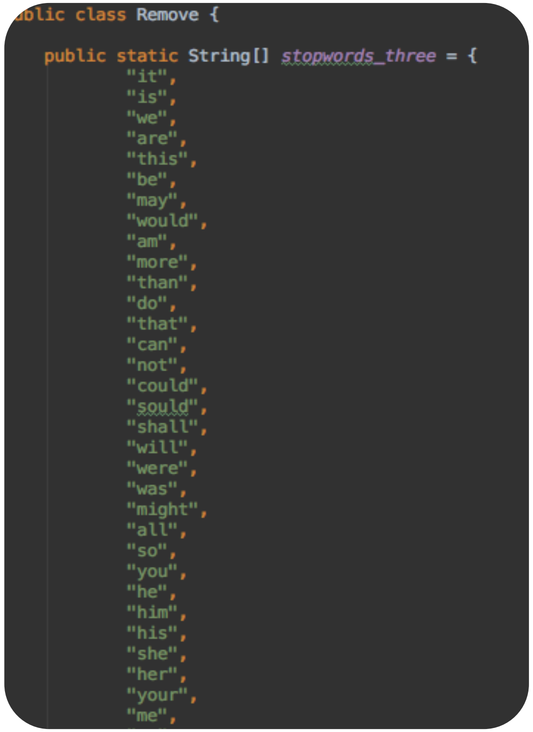


Data processing

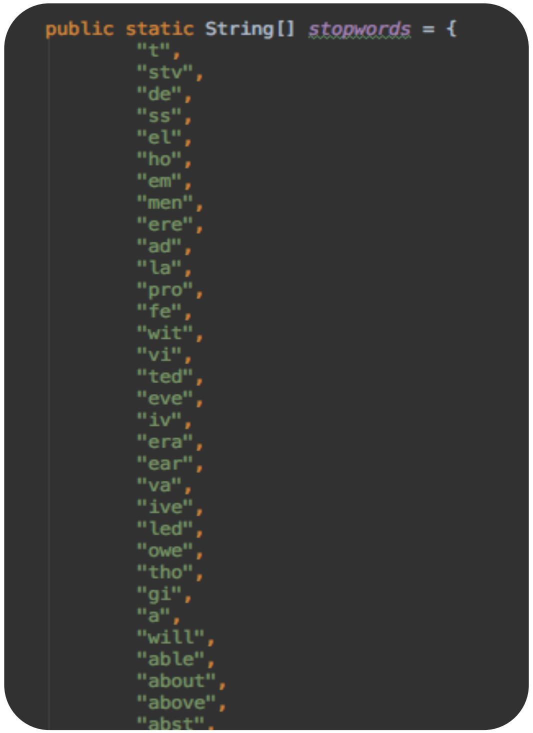


Topic candidate extraction & filtering

- Frequency counting
- Filters :
  - ✓ Stopwords
  - ✓ Thresholds

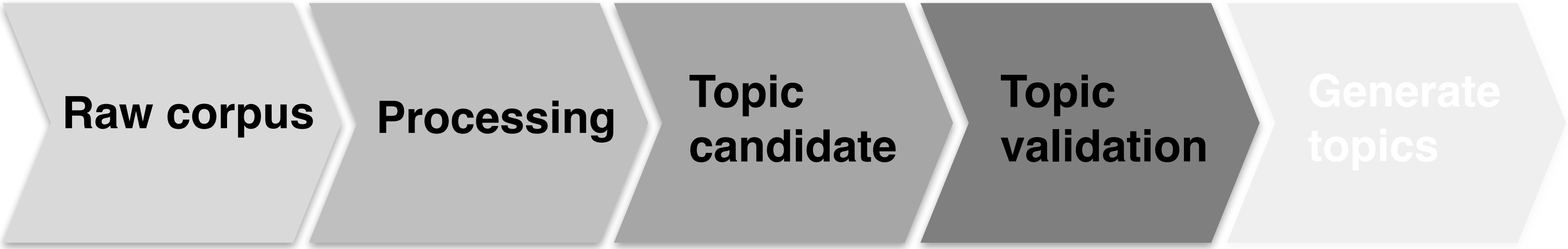


```
ublic class Remove {

    public static String[] stopwords_three = {
            "it",
            "is",
            "we",
            "are",
            "this",
            "be",
            "may",
            "would",
            "am",
            "more",
            "than",
            "do",
            "that",
            "can",
            "not",
            "could",
            "sould",
            "shall",
            "will",
            "were",
            "was",
            "might",
            "all",
            "so",
            "you",
            "he",
            "him",
            "his",
            "she",
            "her",
            "your",
            "me",
```



```
public static String[] stopwords = {
            "t",
            "stv",
            "de",
            "ss",
            "el",
            "ho",
            "em",
            "men",
            "ere",
            "ad",
            "la",
            "pro",
            "fe",
            "wit",
            "vi",
            "ted",
            "eve",
            "iv",
            "era",
            "ear",
            "va",
            "ive",
            "led",
            "owe",
            "tho",
            "gi",
            "a",
            "will",
            "able",
            "about",
            "above",
            "abst",
```

Data processing



Raw corpus

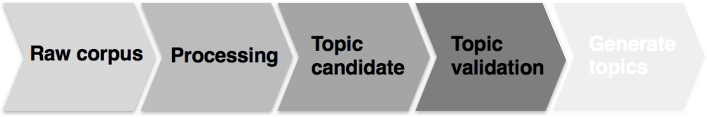
Processing

Topic candidate

Topic validation

Generate topics

Data processing



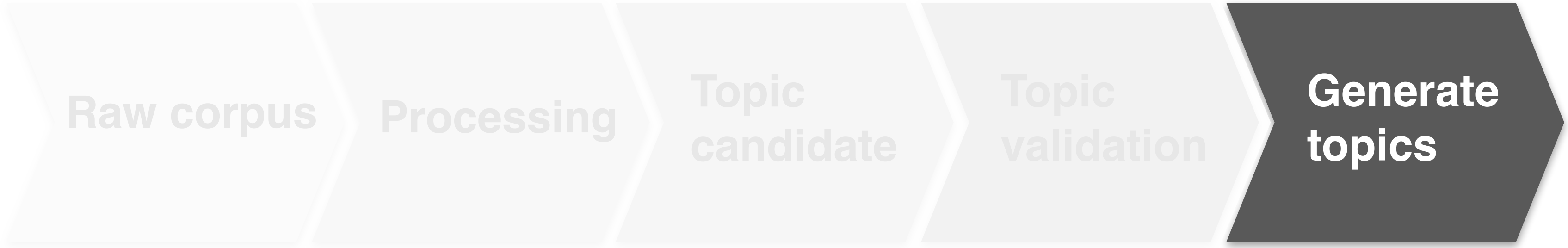
Topic validation

- Human annotation
- Matching
  with **Dictionaries**

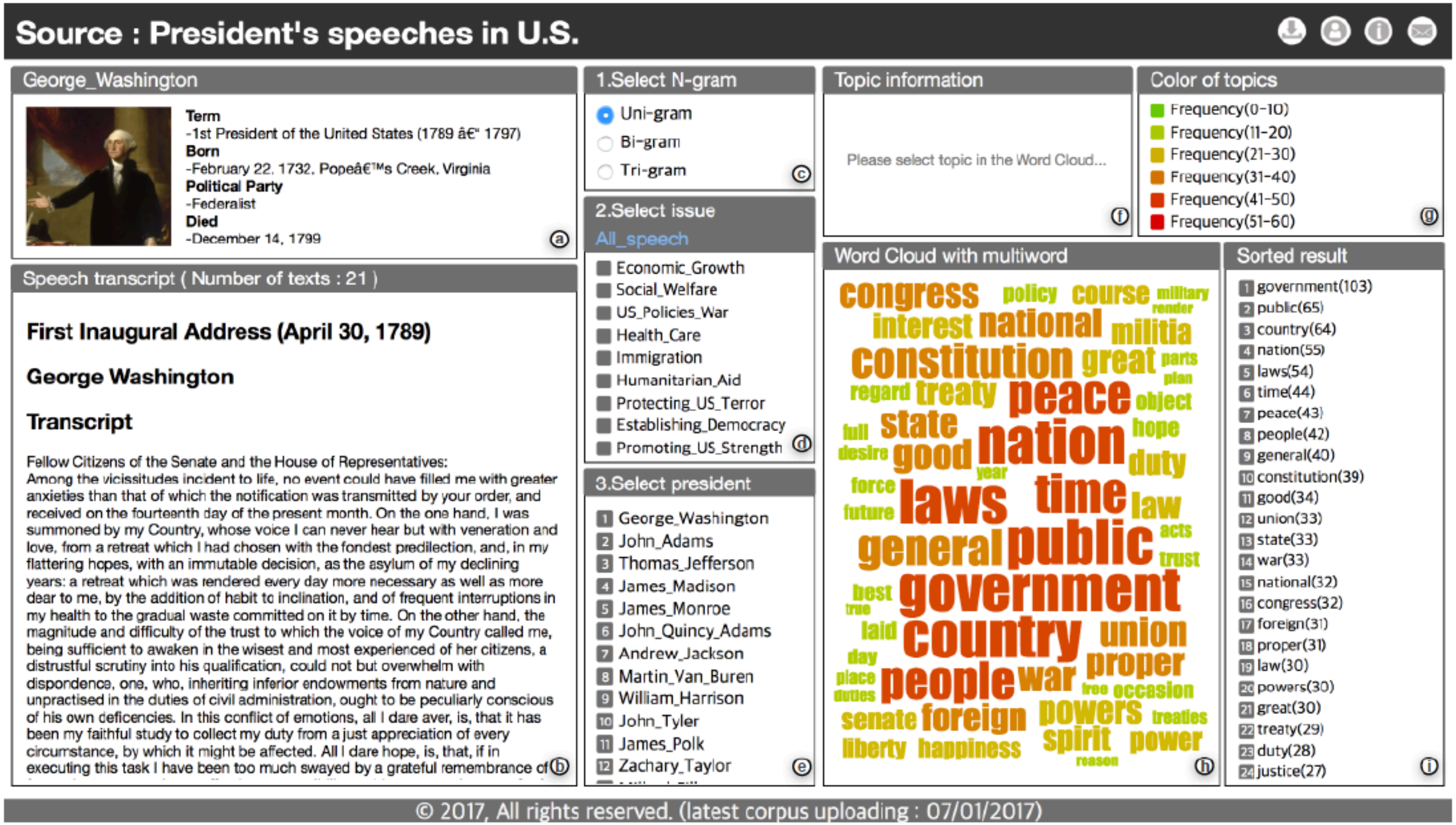
English dictionaries

- THE DEVIL'S DICTIONARY ((C)1911 Released April 15 1993)
- Easton's 1897 Bible Dictionary
- Elements database 20001107
- The Free On-line Dictionary of Computing (27 SEP 03)
- U.S. Gazetteer (1990)
- The Collaborative International Dictionary of English v.0.44
- Hitchcock's Bible Names Dictionary (late 1800's)
- Jargon File (4.3.1, 29 June 2001)
- Virtual Entity of Relevant Acronyms (Version 1.9, June 2002)
- WordNet (r) 2.0
- CIA World Factbook 2002
- User Dictionary

Data processing



Visual system



Source : President's speeches in U.S.

George_Washington

Term
-1st President of the United States (1789 â€" 1797)
Born
-February 22, 1732, Popeâ€™s Creek, Virginia
Political Party
-Federalist
Died
-December 14, 1799

Speech transcript (Number of texts : 21)

First Inaugural Address (April 30, 1789)

George Washington

Transcript

Fellow Citizens of the Senate and the House of Representatives:
Among the vicissitudes incident to life, no event could have filled me with greater anxieties than that of which the notification was transmitted by your order, and received on the fourteenth day of the present month. On the one hand, I was summoned by my Country, whose voice I can never hear but with veneration and love, from a retreat which I had chosen with the fondest predilection, and, in my flattering hopes, with an immutable decision, as the asylum of my declining years: a retreat which was rendered every day more necessary as well as more dear to me, by the addition of habit to inclination, and of frequent interruptions in my health to the gradual waste committed on it by time. On the other hand, the magnitude and difficulty of the trust to which the voice of my Country called me, being sufficient to awaken in the wisest and most experienced of her citizens, a distrustful scrutiny into his qualification, could not but overwhelm with dispondence, one, who, inheriting inferior endowments from nature and unpractised in the duties of civil administration, ought to be peculiarly conscious of his own deficencies. In this conflict of emotions, all I dare aver, is, that it has been my faithful study to collect my duty from a just appreciation of every circumstance, by which it might be affected. All I dare hope, is, that, if in executing this task I have been too much swayed by a grateful remembrance of

1.Select N-gram

● Uni-gram
○ Bi-gram
○ Tri-gram

2.Select issue

All_speech

- ■ Economic_Growth
- ■ Social_Welfare
- ■ US_Policies_War
- ■ Health_Care
- ■ Immigration
- ■ Humanitarian_Aid
- ■ Protecting_US_Terror
- ■ Establishing_Democracy
- ■ Promoting_US_Strength

3.Select president

1 George_Washington
2 John_Adams
3 Thomas_Jefferson
4 James_Madison
5 James_Monroe
6 John_Quincy_Adams
7 Andrew_Jackson
8 Martin_Van_Buren
9 William_Harrison
10 John_Tyler
11 James_Polk
12 Zachary_Taylor

Topic information

Please select topic in the Word Cloud...

Color of topics

- 🟩 Frequency(0-10)
- 🟩 Frequency(11-20)
- 🟨 Frequency(21-30)
- 🟧 Frequency(31-40)
- 🟧 Frequency(41-50)
- 🟥 Frequency(51-60)

Word Cloud with multiword

Sorted result

1	government(103)
2	public(65)
3	country(64)
4	nation(55)
5	laws(54)
6	time(44)
7	peace(43)
8	people(42)
9	general(40)
10	constitution(39)
11	good(34)
12	union(33)
13	state(33)
14	war(33)
15	national(32)
16	congress(32)
17	foreign(31)
18	proper(31)
19	law(30)
20	powers(30)
21	great(30)
22	treaty(29)
23	duty(28)
24	justice(27)

© 2017, All rights reserved. (latest corpus uploading : 07/01/2017)

http://ressources.modyco.fr/sm/MultiwordVis/

Université
Paris Nanterre

AJOU UNIVERSITY

Ambiguous sentence

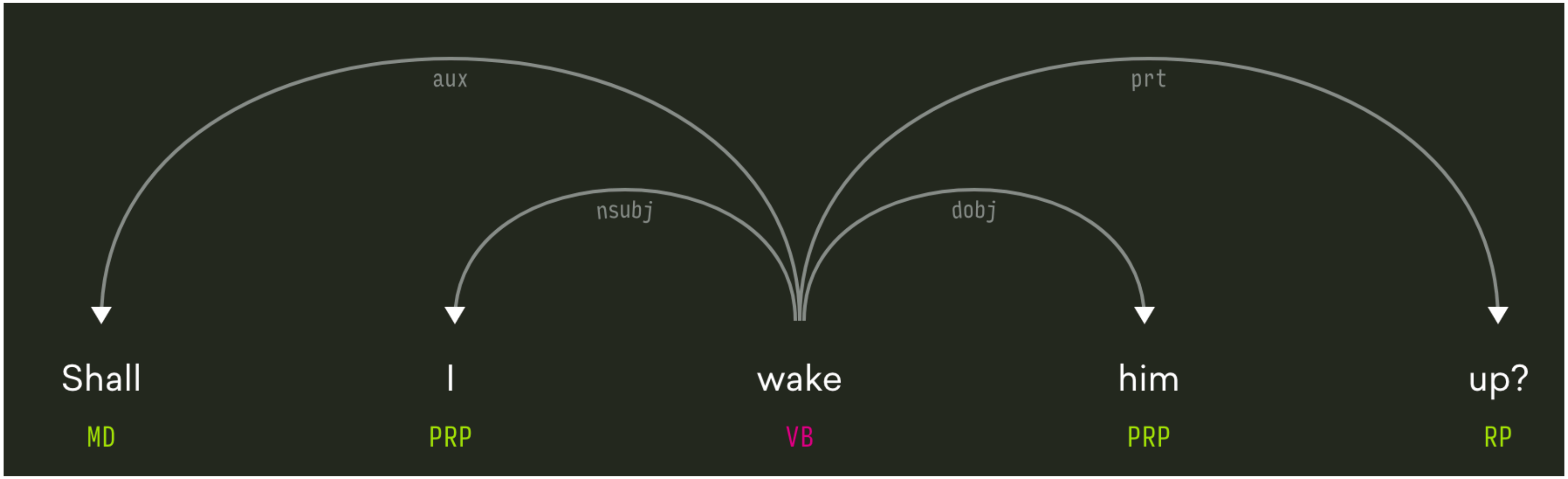
“Shall I wake him up?”

Ambiguous sentence

“Shall I **wake** him **up**?”

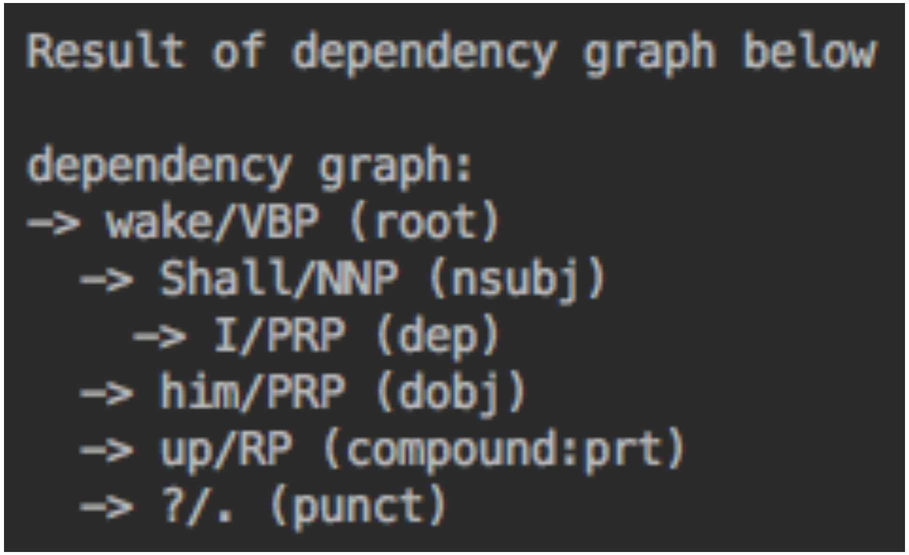
We can’t extract wake up if we only use N-gram algorithm.

Dependency tag

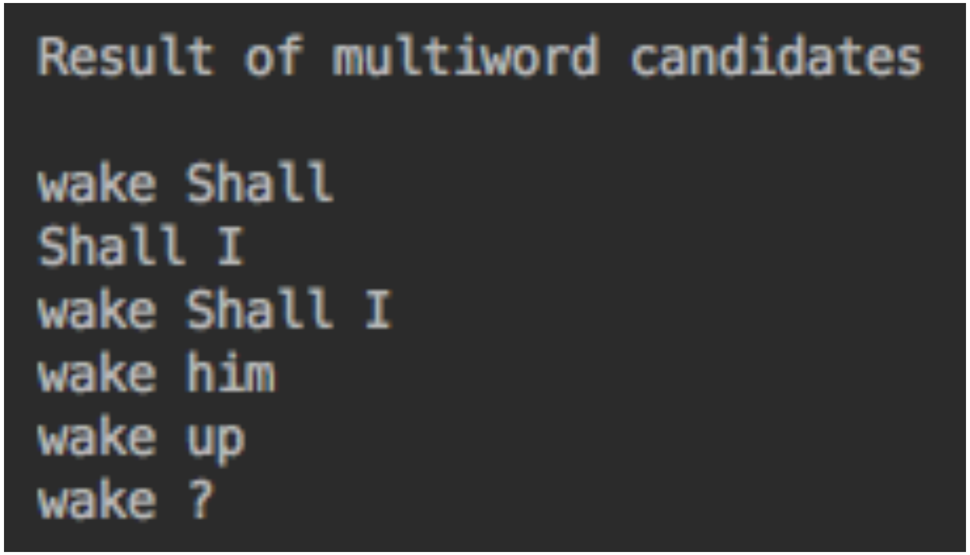


Dependency tag can provide a simple description of the grammatical relationships in a sentence.

Improving algorithm

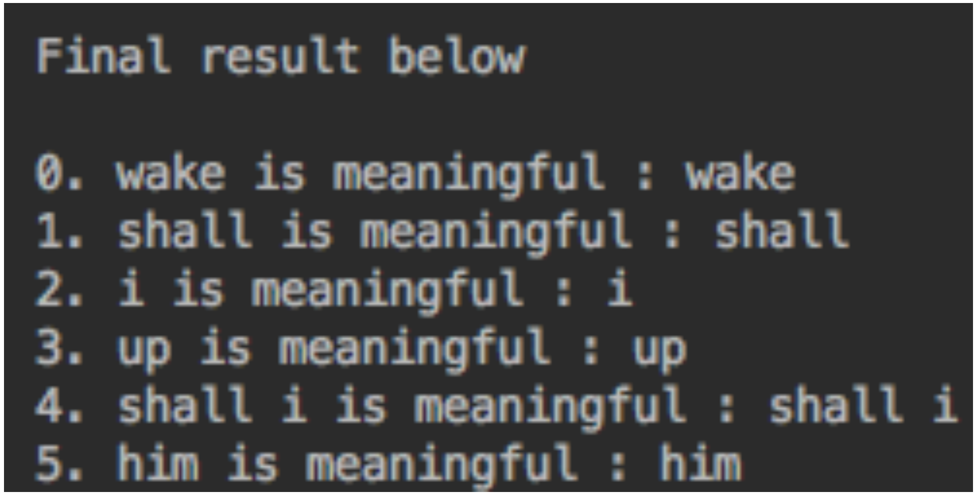


```
Result of dependency graph below

dependency graph:
-> wake/VBP (root)
  -> Shall/NNP (nsubj)
    -> I/PRP (dep)
  -> him/PRP (dobj)
  -> up/RP (compound:prt)
  -> ?/. (punct)
```



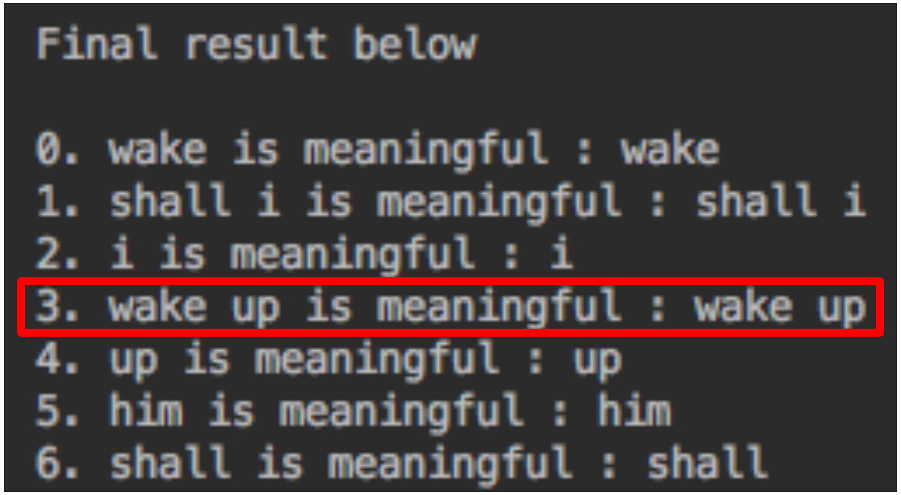
```
Result of multiword candidates

wake Shall
Shall I
wake Shall I
wake him
wake up
wake ?
```

Improving algorithm



```
Final result below

0. wake is meaningful : wake
1. shall is meaningful : shall
2. i is meaningful : i
3. up is meaningful : up
4. shall i is meaningful : shall i
5. him is meaningful : him
```

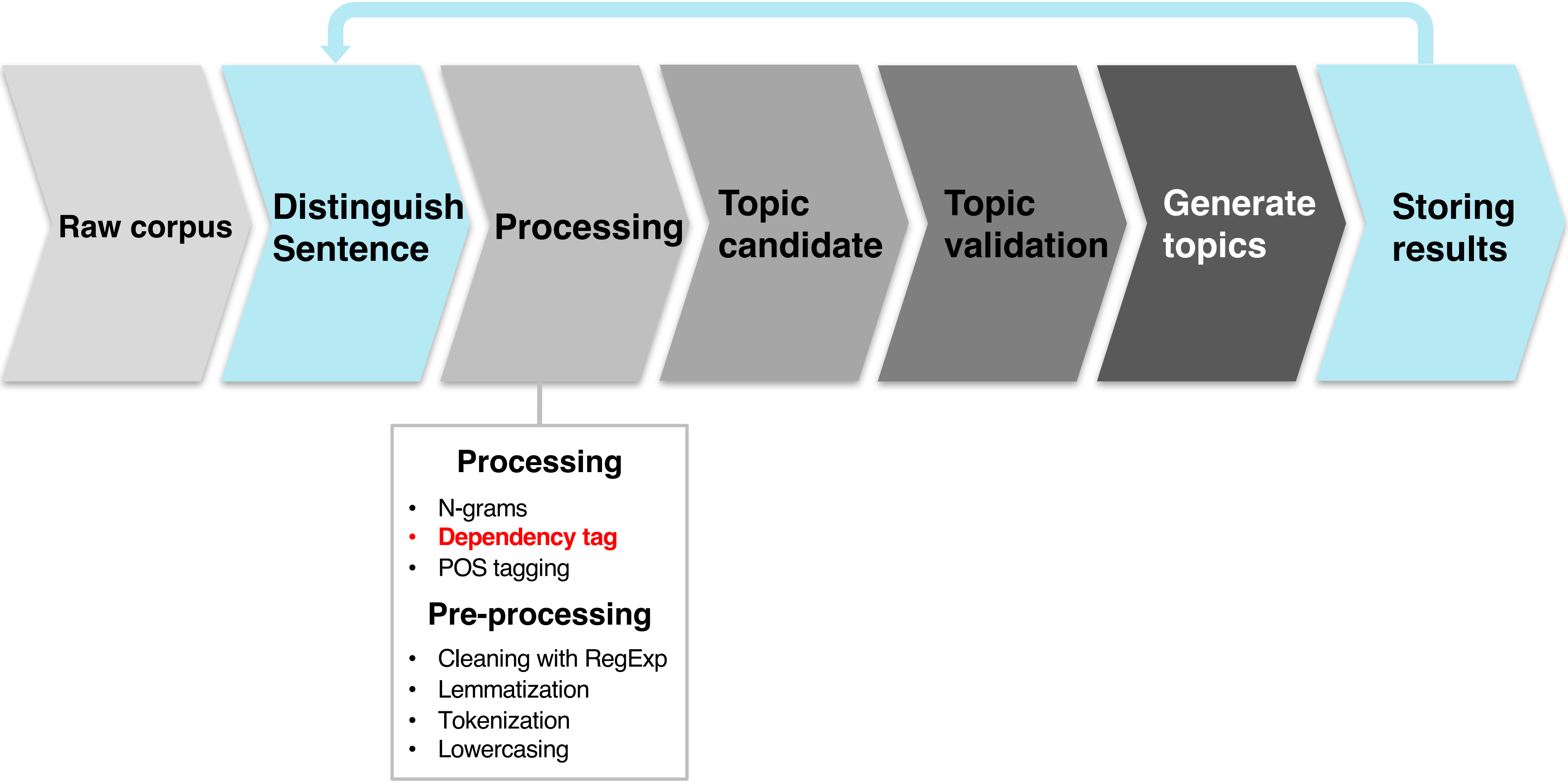
N-gram



```
Final result below

0. wake is meaningful : wake
1. shall i is meaningful : shall i
2. i is meaningful : i
3. wake up is meaningful : wake up
4. up is meaningful : up
5. him is meaningful : him
6. shall is meaningful : shall
```

Dependency tag

Data processing



Processing

- N-grams
- Dependency tag
- POS tagging

Pre-processing

- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

Université Paris Nanterre AJOU UNIVERSITY

Q&A

Thank you for listening.

stat34@ajou.ac.kr

Université Paris Nanterre AJOU UNIVERSITY