


Castro Speeches Summarization tool

A large crowd of people gathered for a speech, with a man in a green suit addressing them. The background shows a cityscape with several buildings under a clear sky.

The In-Fidels

Ervina Cergani, Benedict Fehringer, Besnik Fetahu, Evangelia Kiagia, Christian Wellner

Outline

Background

- ❖ Data Set
- ❖ Pipeline

Predefined Queries

- ❖ Preprocessing
- ❖ IR & IE methods

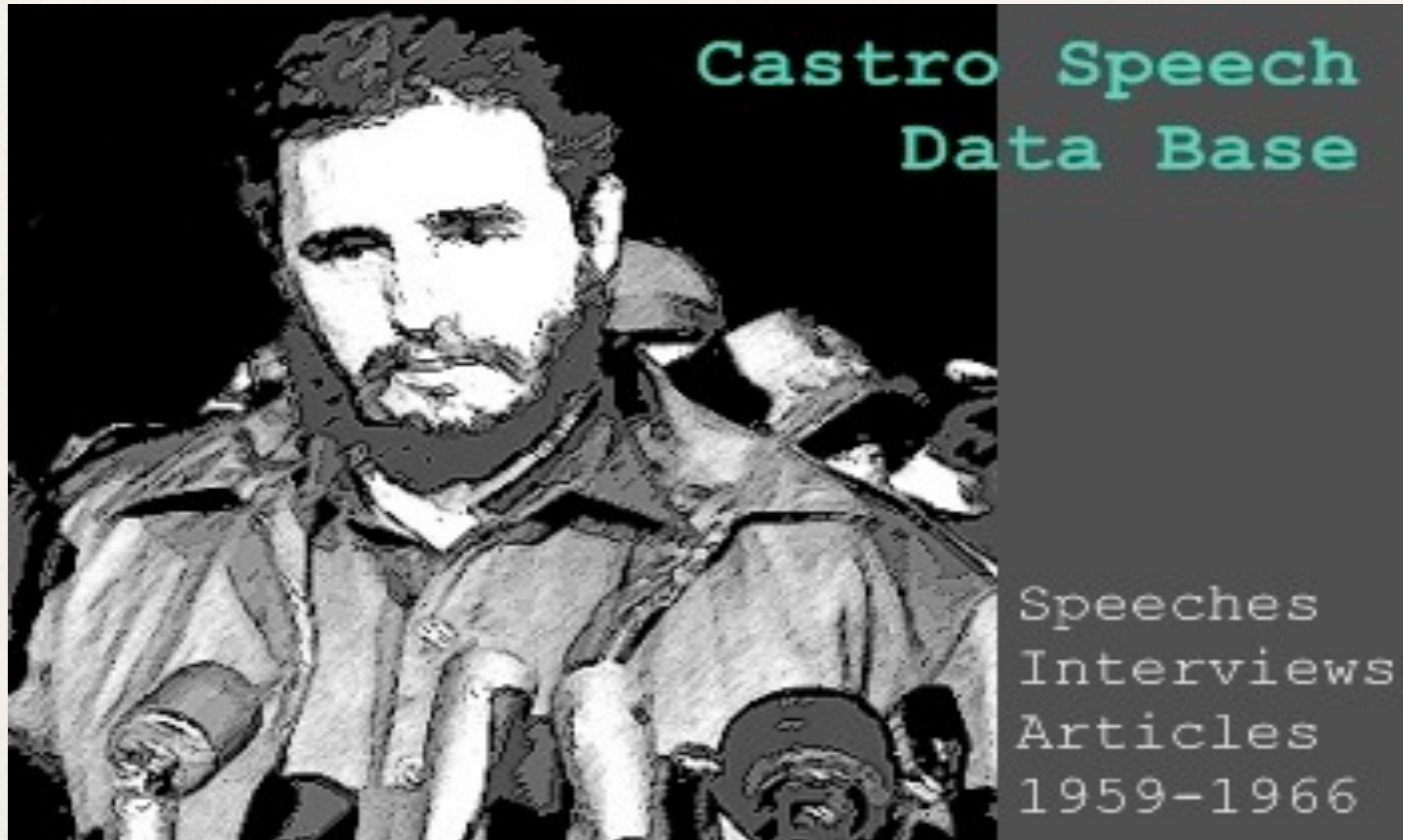
Non-defined Queries

Demonstration

Problems

Future Directions

Data Set



Data set provided by:LANIC

- ❖ Latin American Network Information Center(LANIC)
- ❖ Affiliated and supported by the Institute of Latin American Studies of the University of Texas at Austin.
- ❖ Lanic hosts extensive digital content on Latin America through its Collection

“Castro Speech Database”

- ❖ 2,000 interviews, speeches and press conferences, from 1959-1996
- ❖ Structured data
- ❖ Based on the records of Foreign Broadcast Information(FBIS)

```
-DATE-  
19641101  
-YEAR-  
1964  
-DOCUMENT_TYPE-  
SPEECH  
-AUTHOR-  
F. CASTRO  
-HEADLINE-  
CASTRO SEES PROGRESS IN REVOLUTIONARY SPIRIT  
-PLACE-  
CTC-R THEATER IN HAVANA  
-SOURCE-  
HAVANA DOMESTIC RADIO  
-REPORT_NBR-  
FBIS  
-REPORT_DATE-  
19641103  
-TEXT-  
CASTRO SEES PROGRESS IN REVOLUTIONARY SPIRIT
```

Focus Topics

- ❖ Social Development
- ❖ Education
- ❖ Medical Domain
- ❖ Agriculture
- ❖ Industry
- ❖ Why:
 - ❖ Social development goes hand in hand with the historical, political and ideological situation of the time.

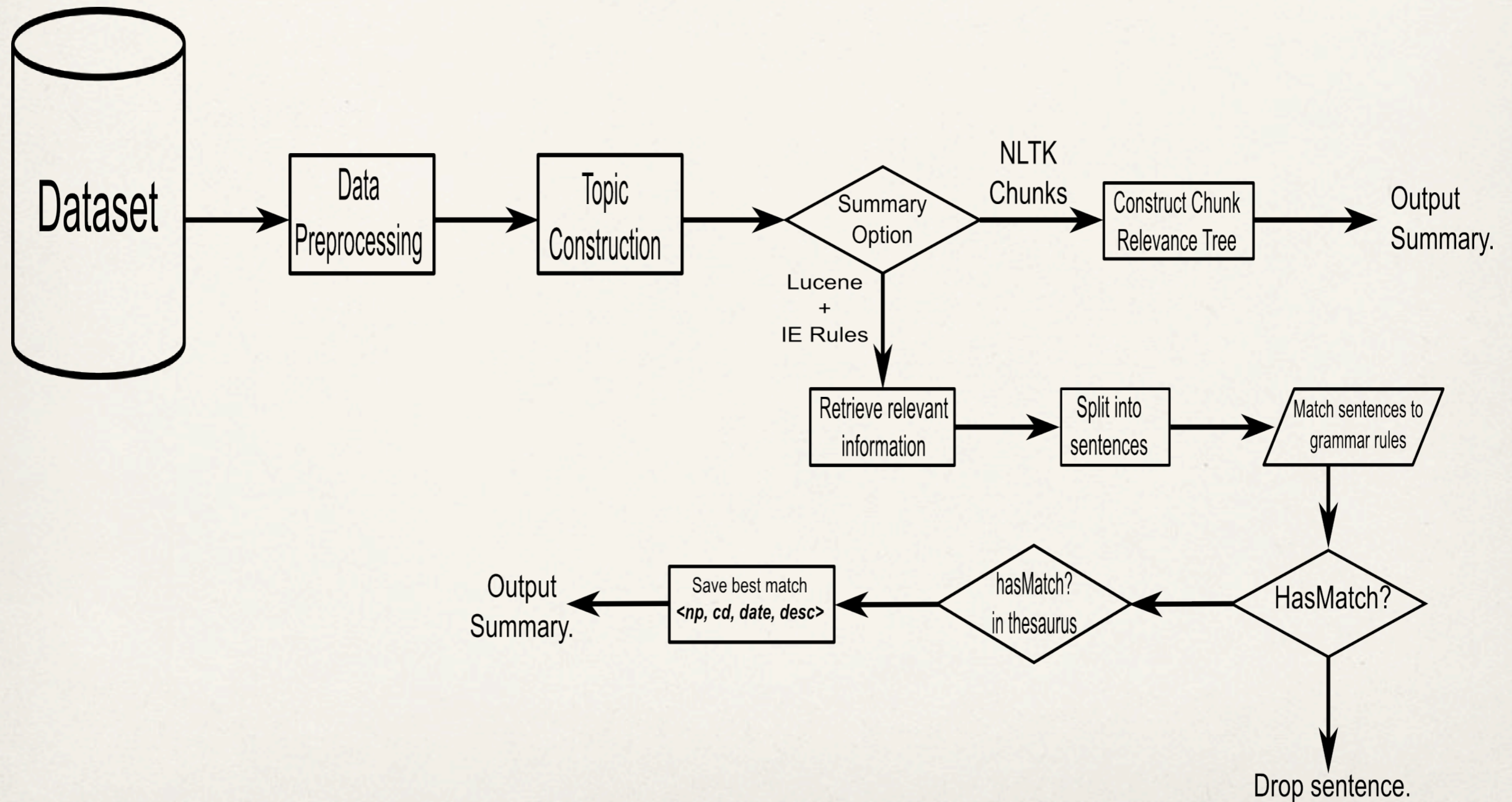
Historical Facts in data set:

- ❖ Progress of Cuban social development throughout 4 decades (50's-90's)
- ❖ Gradual nationalization of the industry
- ❖ External Policy and relations
- ❖ Grains of communist ideology and its impact to internal and external policy
- ❖ Phases of the economic embargo imposed by the different US presidents and governments

Goal

- ❖ Direct accessible primary sources for historical research
- ❖ Categorized & summarized statistical data of actual facts and figures of Cuban social development(predefined-query)
- ❖ Search and query for specific terms through non-defined query
- ❖ Parallel access to the original speeches for double checking

Pipeline



Predefined Queries

Data Pre-processing

Document File

The interest with which this congress has been received and the high degree of participation to which we have referred constitute a concrete expression of the community of interests being increasingly manifested among the workers of the world. We can point out that the congress we are pleased to inaugurate today is the congress of the large majority of the organized trade union movement on an international level.

This is the first time that a congress of this type has been held outside of Europe and the fact that it is being held in Cuba -- a country which is struggling for development in the midst of continued threats, slanderous campaigns and a rigorous economic blockade; a country whose enemies have tried to isolate and banish from the rest of the peoples of the world - gives it a solidary significance that our people and government profoundly appreciate.

We, of course, do not ignore the heterogeneity of the political, philosophical and religious positions assembled in this hall, the existing diversity of opinions among many of the organizations present here; and the fact that this congress has the peculiarity of assembling union leaders from socialist countries, union leaders who are active in capitalist countries, labor leaders active in highly industrialized states and leaders from the largely underdeveloped and economically backward areas of the world. There is a great variety of circumstances and opinions. Under such conditions, could it be possible to find a common language? We believe it is possible, that it is necessary and, still more, indispensable.

RegEx

Paragraph Files

The interest with which this congress has been received and the high degree of participation to which we have referred constitute a concrete expression of the community of interests being increasingly manifested among the workers of the world.

This is the first time that a congress of this type has been held outside of Europe and the fact that it is being held in Cuba - a country which is struggling for development in the midst of continued threats, slanderous campaigns and a rigorous economic blockade

• • • • •

We, of course, do not ignore the heterogeneity of the political, philosophical and religious positions assembled in this hall, the existing diversity of opinions among many of the organizations present here; and the fact that this congress has the peculiarity of assembling union leaders.

NLTK

POS - Paragraphs Files

The_DT interest_NN with_IN which_WDT this_DT congress_NN has_VBZ been_VBN received_VBN and_CC the_DT high_JJ degree_NN of_IN participation_NN to_TO which_WDT we_PRP have_VBP referred_VBN constitute_NN a_DT concrete_JJ expression_NN of_IN

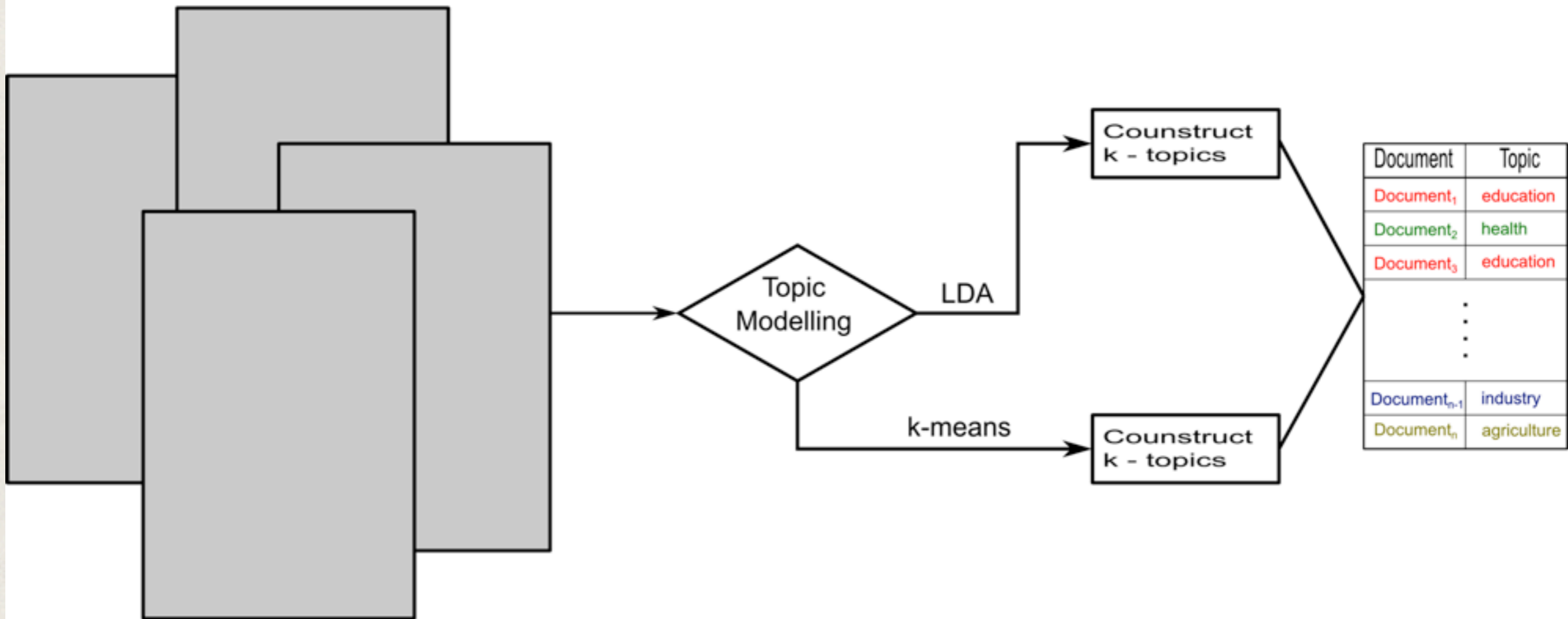
This_DT is_VBZ the_DT first_JJ time_NN that_WDT a_DT congress_NN of_IN this_DT type_NN has_VBZ been_VBN held_VBN outside_IN of_IN Europe_NNP and_CC the_DT fact_NN that_IN it_PRP is_VBZ being_VBG held_VBN in_IN Cuba_NNP a_DT country_NN which_WDT

• • • • •

We_PRP , of_IN course_NN , do_VBP not_RB ignore_VB the_DT heterogeneity_NN of_IN the_DT political_JJ , philosophical_JJ and_CC religious_JJ positions_NNS assembled_VBN in_IN this_DT hall_NN , the_DT existing_VBG diversity_NN of_IN

Document Categorization

Documents



LDA Topic Modelling Tool

- ❖ “Latent Dirichlet Allocation”
- ❖ Generative probabilistic model for collection of data
- ❖ Each document is a random mixture of corpus-wide topics
- ❖ Each word is drawn from these topics

LDA Topic Modelling Tool

- ❖ LDA puts “topical” words together
- ❖ Word probabilities are maximized by dividing the words among the topics.

```
Console
<100> LL/token: -8.76584
<110> LL/token: -8.75791
<120> LL/token: -8.74284
<130> LL/token: -8.75137
<140> LL/token: -8.75303
0      10      development year road percent part grade revolution day things words
1      10      school education million schools future years center peasants thousands level
2      10      area sugarcane sugar people children time plan rice water cane
3      10      country applause teachers plans planted pupils areas services polyclinic persons
4      10      land production students work room consumption increase roads peasant number

<150> LL/token: -8.74535
<160> LL/token: -8.7518
<170> LL/token: -8.75169
<180> LL/token: -8.76955
<190> LL/token: -8.77968
0      10      development year teachers planted percent grade revolution part increase town
1      10      school education million time future road years center persons things
2      10      country sugarcane area sugar people children areas plan rice peasant
3      10      applause production plans pupils services day roads province el shouts
4      10      schools land students work cane room consumption conditions polyclinic present

<200> LL/token: -8.77418

Total time: 1 seconds
Mallet Output files written in /home/evi/Downloads --> /home/evi/Downloads/output_state.gz . /home/evi/Downloads/output_topic_keys

Csv Output files written in /home/evi/Downloads/output_csv
Html Output files written in /home/evi/Downloads/output_html

PROCESS COMPLETE
Time :6.437
```

Blei.D. (2009) “Topic Models” Computer Science Department,
Princeton University

http://videolectures.net/mlss09uk_blei_tm/

Document Categorization - I

tf-idf for comparison.

- ❖ Cluster documents, using k-means.
- ❖ Tanimoto distance, for document similarity.

$$f(A, B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$

$$A \cdot B = \sum_i (A_i \wedge B_i) \quad |A|^2 = \sum_i (A_i)$$

Document Categorization - I

- * Cluster documents, using k-means.
- * Tanimoto distance, for document similarity.

$$f(A, B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$

$$A \cdot B = \sum_i (A_i \wedge B_i) \quad |A|^2 = \sum_i (A_i)$$

tf-idf for comparison.

Document	19590103
-----	-----
cantillo	0.03271093930539276
santiago	0.024037028819946824
military	0.020506722190177733
rego	0.01833223516715711
batista	0.014500865040818081
coup	0.01424331669904718
colonel	0.014101872594346063
republic	0.014037788432747023
colombia	0.01209614894565306
urrutia	0.012085910953024004
piedra	0.011890818179504653
uprising	0.011372718162631608
garrison	0.011303225966340738
-----	-----

Document Categorization - I

- * Cluster documents, using k-means.
- * Tanimoto distance, for document similarity.

$$f(A, B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$
$$A \cdot B = \sum_i (A_i \wedge B_i) \quad |A|^2 = \sum_i (A_i)$$

tf-idf for comparison.

Document	19590103
canti	
santi	
milit	
rego	
batis	
coup	
colon	
repub	
colom	
urru	
piedr	
upris	
garri	

duty	

Document	19590109-1
peace	0.026927216822081632
weapons	0.02164912406813786
army	0.01897175376127404
republic	0.017527101158547692
revolutionaries	0.014883394830339243
ambition	0.014367622595955983
command	0.013430331653104356
rebel	0.013107237258996804
merit	0.011867782182669795
guns	0.011608839986122712
fought	0.011145819973216523
duty	0.010724390184021722

Main thematic centers of speeches

- ❖ Education
- ❖ Medical Domain
- ❖ Agriculture
- ❖ Industry
- ❖ Ideology
- ❖ External Policy

Retrieval using Lucene

- ❖ Query operations using lucene:

- ❖ Full text,

- ❖ Date ranges,

- ❖ Indexing of data files, etc.

- ❖ Sample query:

TEXT: health institutions **DATE:** [19800101 TO 19900101] **CATEGORY:** Health

IE: Grammar Patterns

- ❖ Grammatical patterns that capture statistical data

“ By 1985 we will have nearly 20,000 physicians ”

- ❖ Patterns that capture dates

“In 1970 the production of sugar cane has doubled”

- ❖ Patterns that capture numbers

“We had constructed more than 3,000 farms”

Specific Rules:

- ❖ **IN+CD+RPR+MD+VB+RB
+CD+NNS**
- ❖ **IN+CD+DT+NN+IN+NN+NN
+VB+VBD**
- ❖ **PRP+VB+VBD+JJR+IN+CD**

Evaluation Results

Category	Summary Objects	Correct	Not Correct	Precision
Education	pupils, schools, students, teachers, statesman, specialist, technician, scholarship, triumph, professor, job	77	43	0.64
Health	home	1	0	1.0
Industry	wage, growth	1	1	0.5
Agriculture	cane, mill, trucks, ranch, petroleum, plant, rice	18	3	0.85

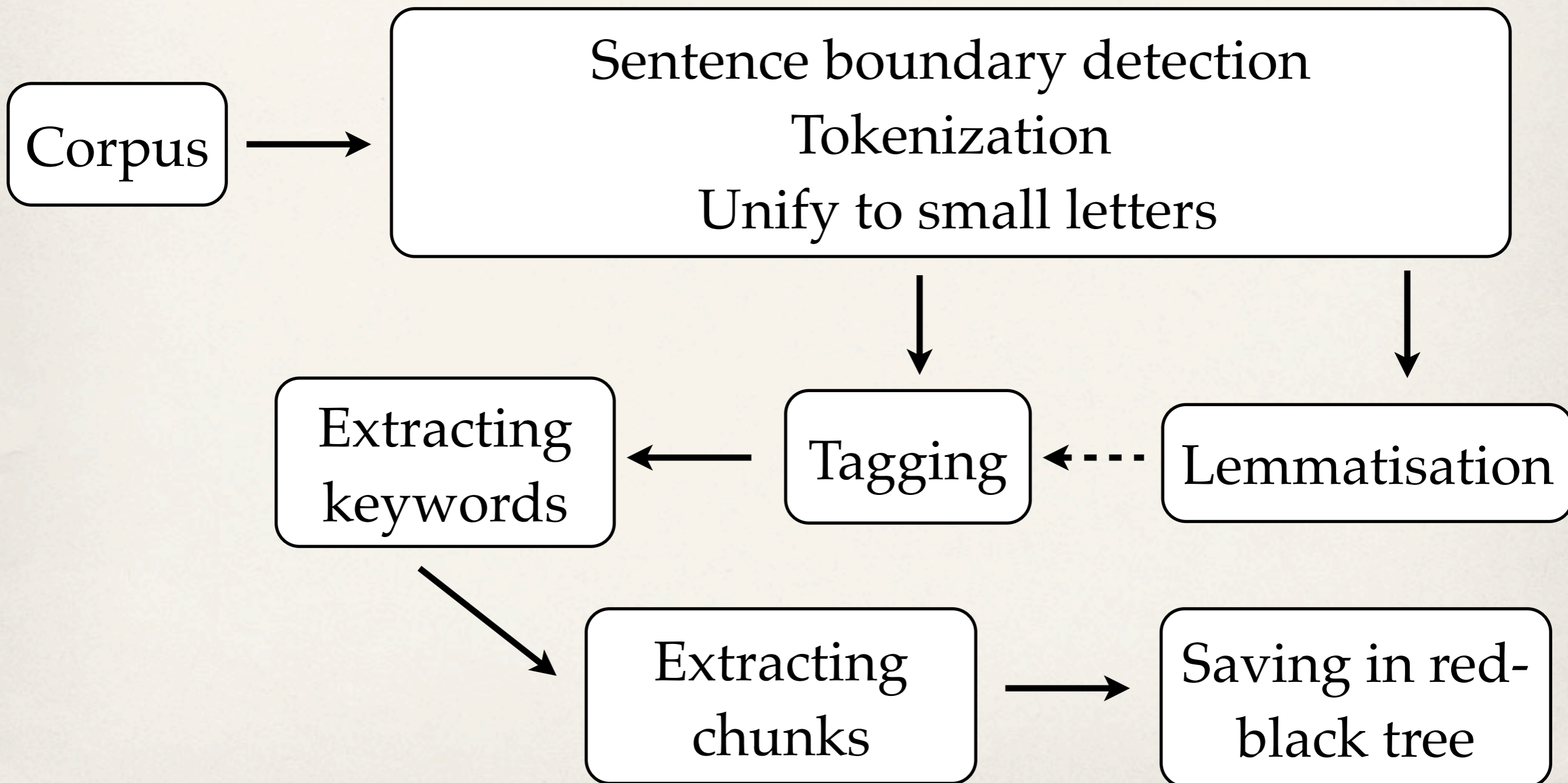
Non-defined Queries

Goal

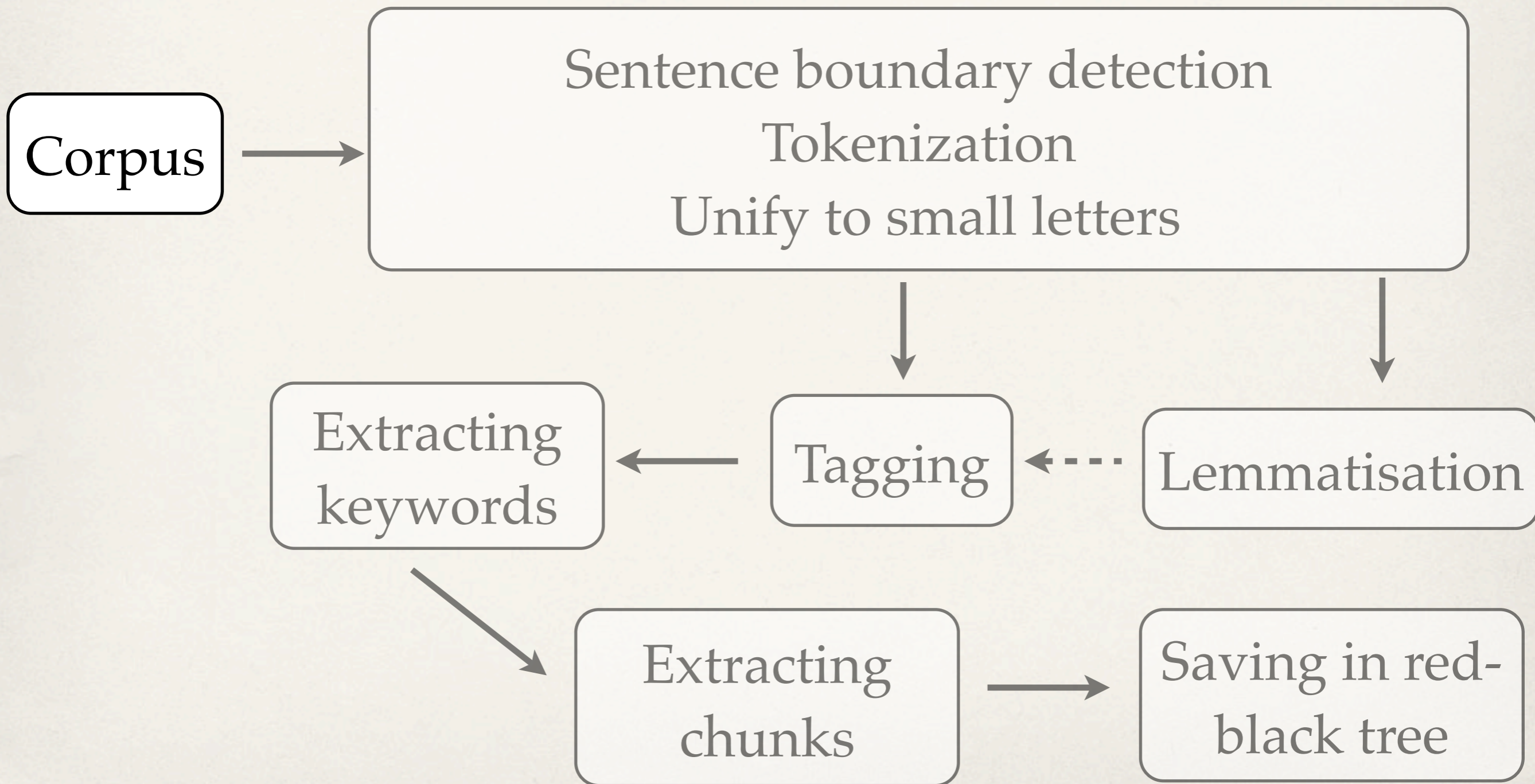
Goal

- ❖ Provide information for **every query**
 - ❖ sufficient **much** information to be **informative**
 - ❖ sufficient **little** information to **deal with**

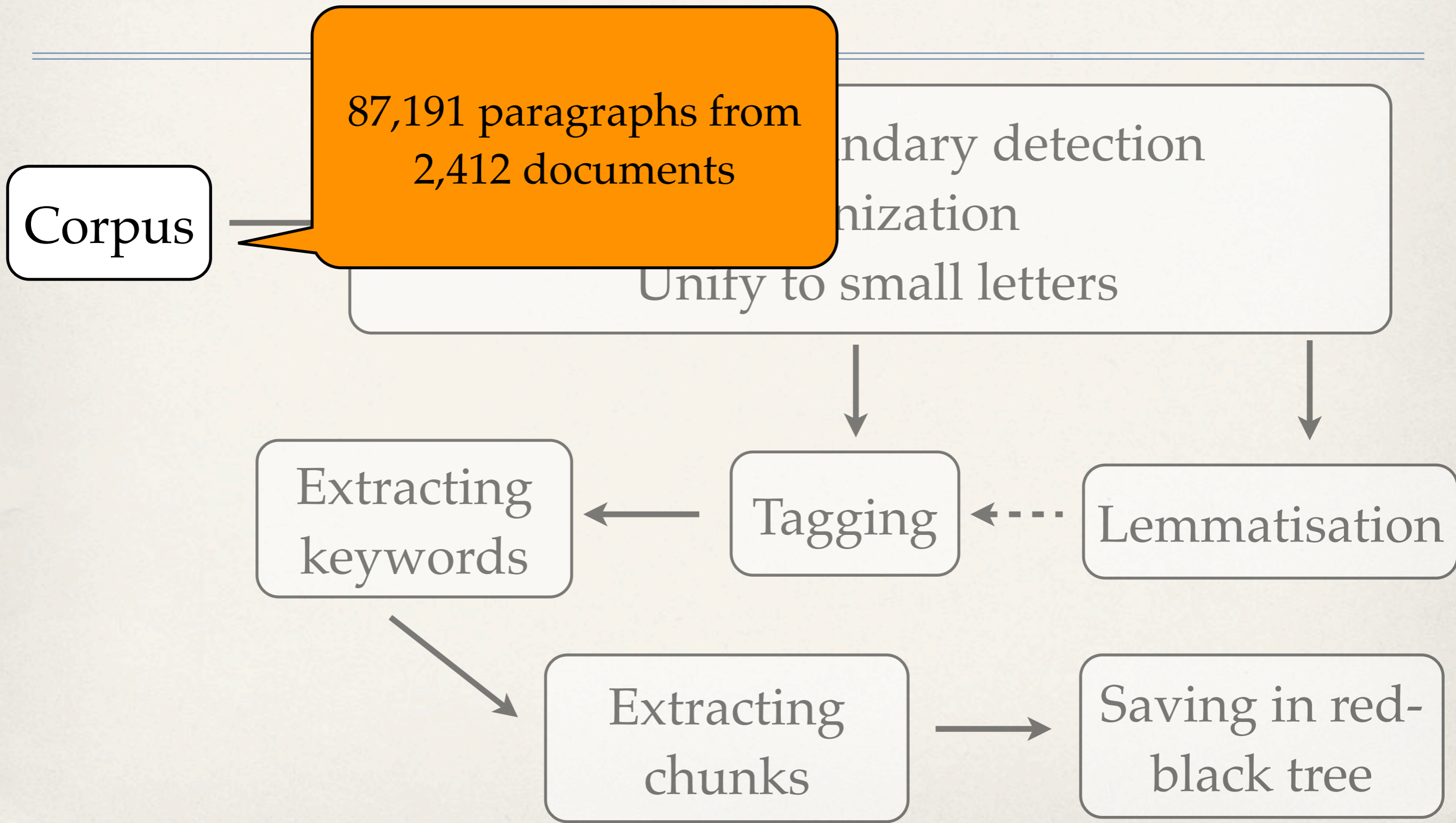
Preprocessing



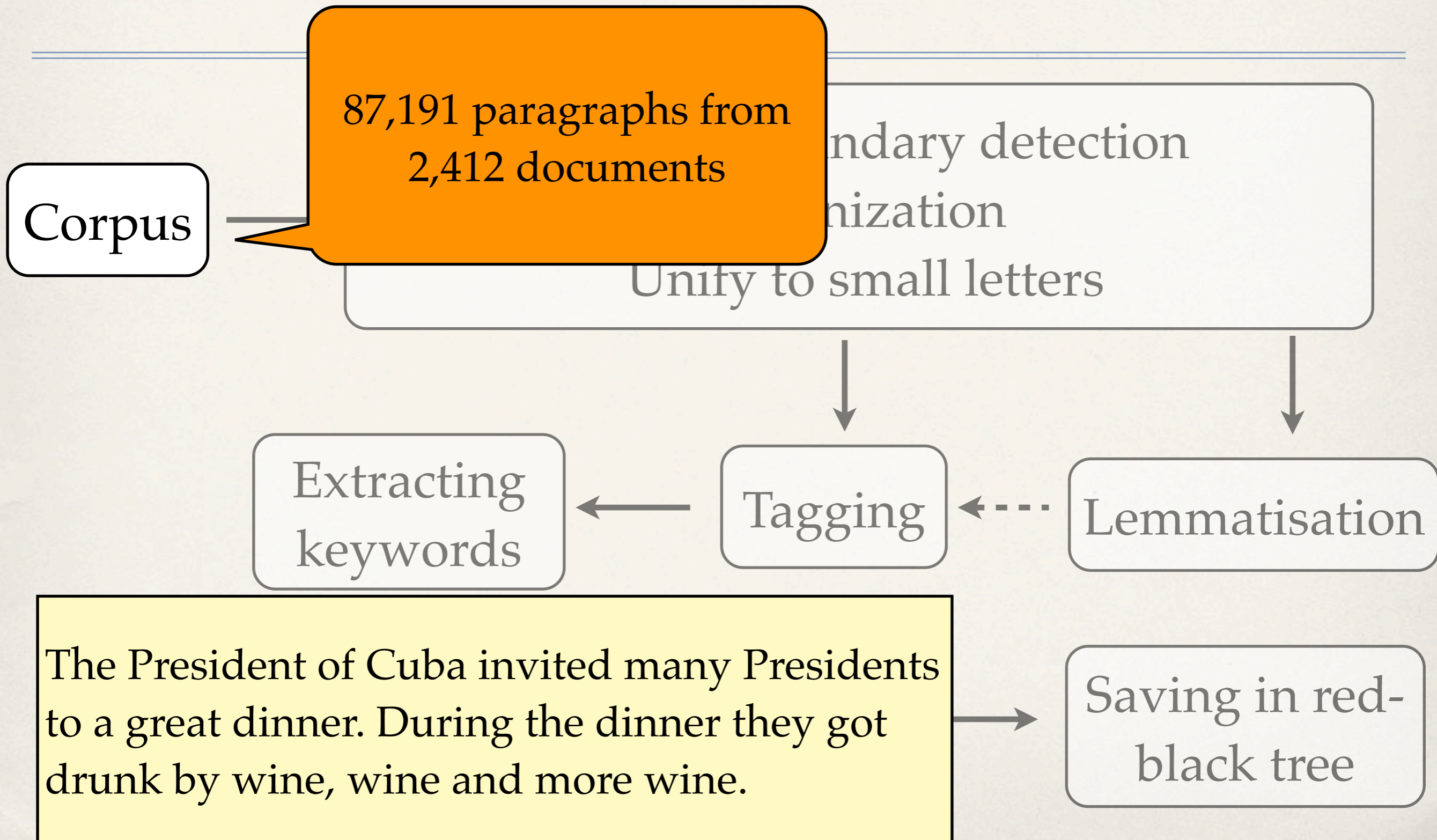
Preprocessing I



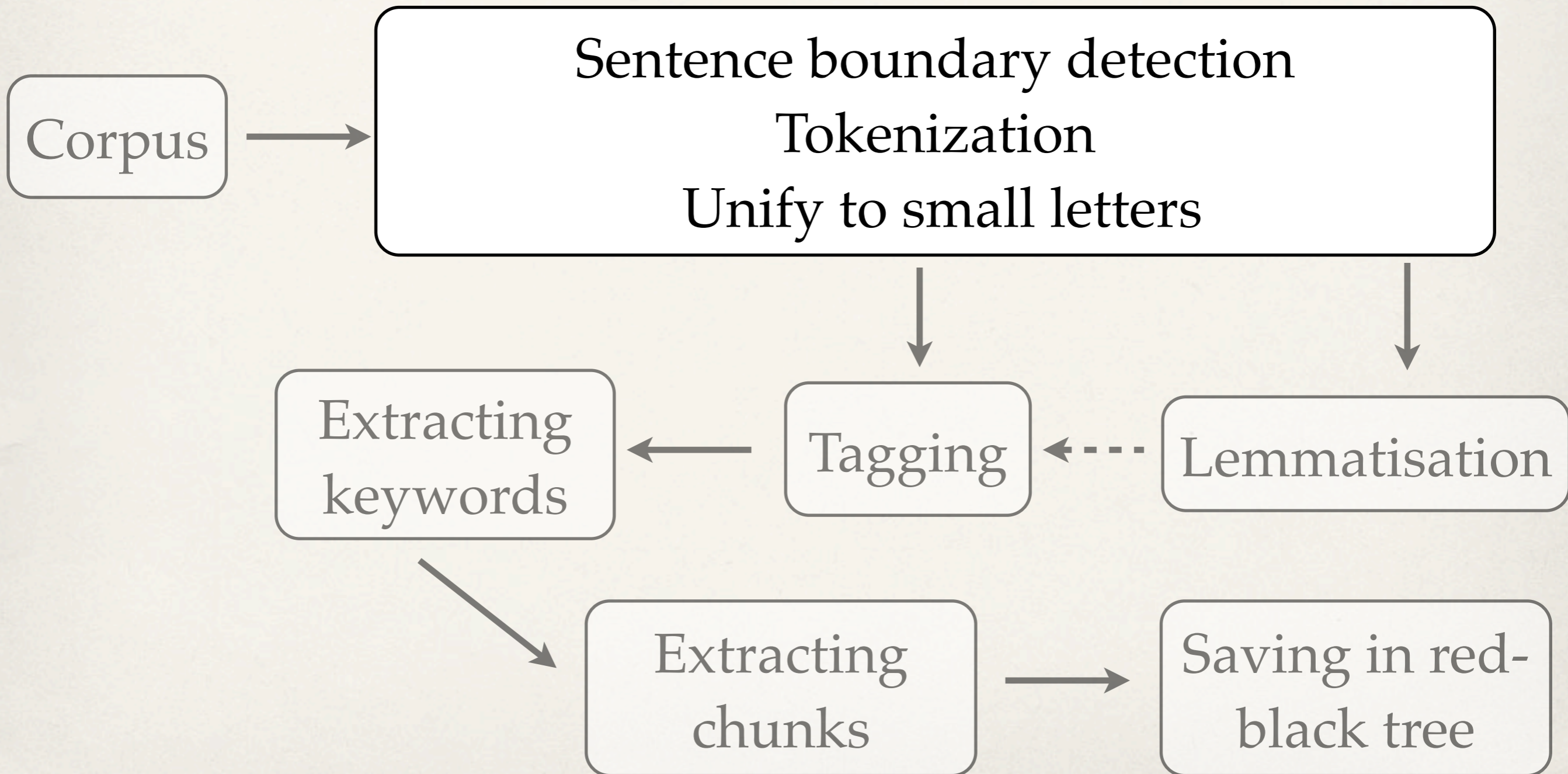
Preprocessing I



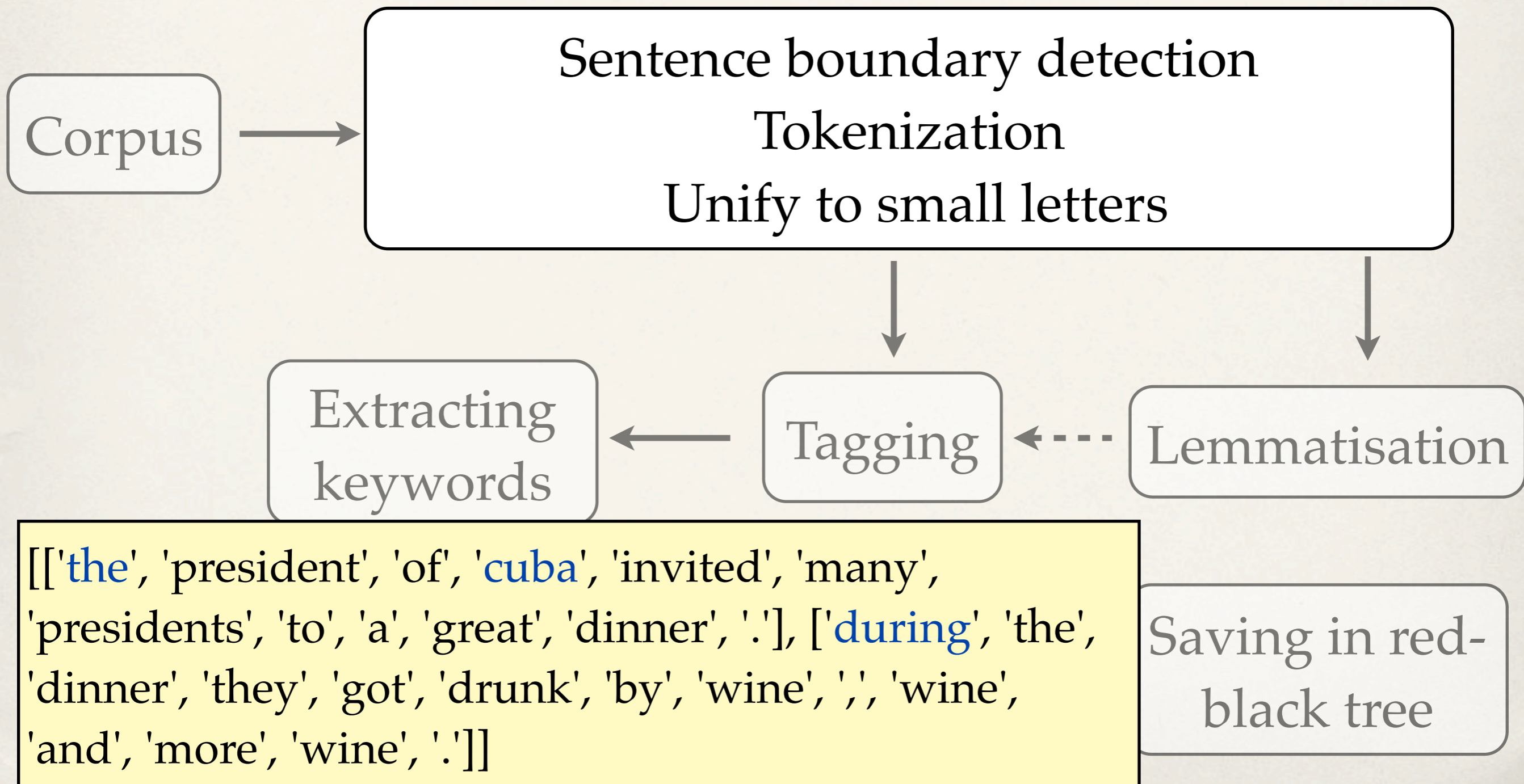
Preprocessing I



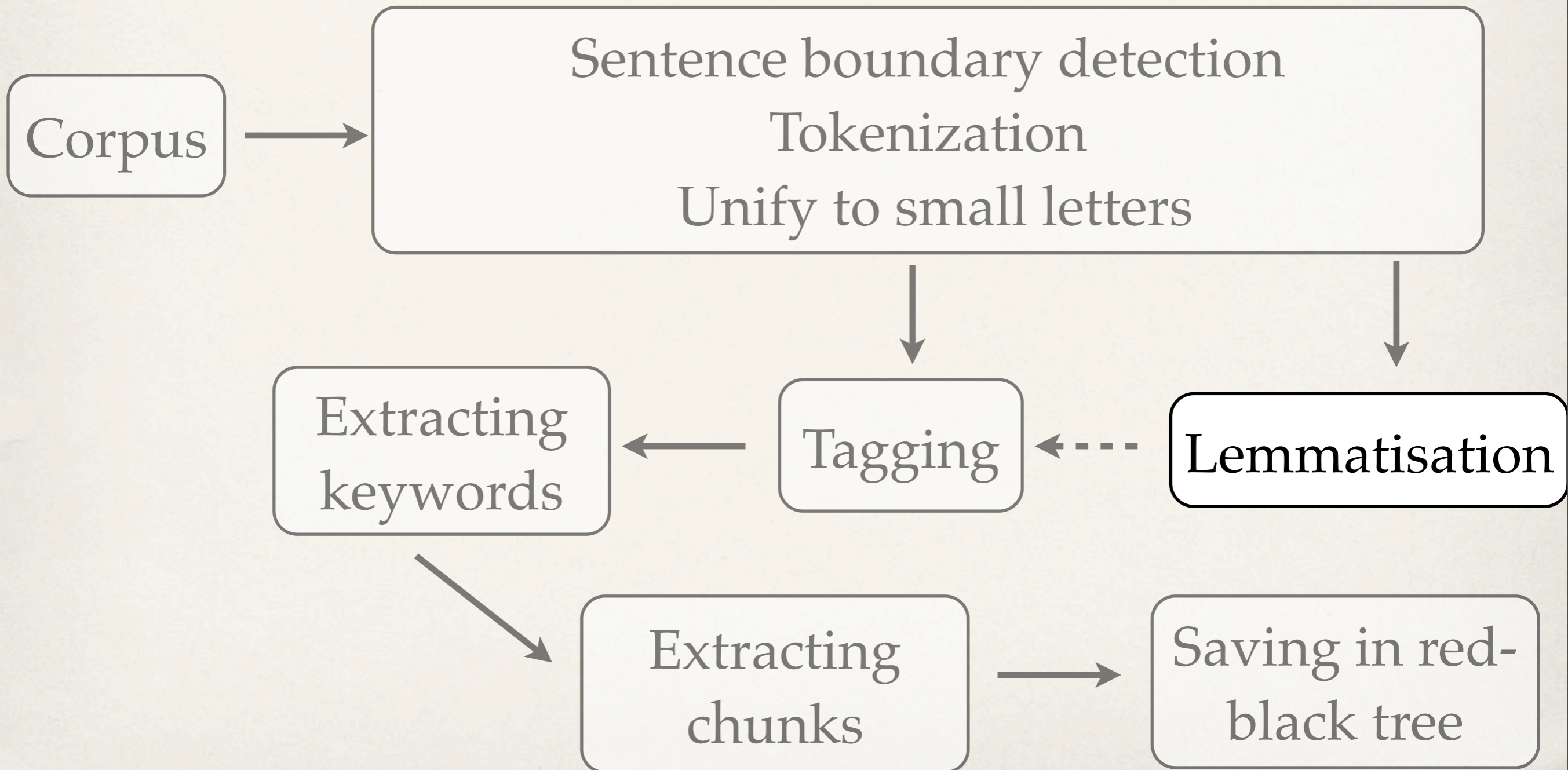
Preprocessing



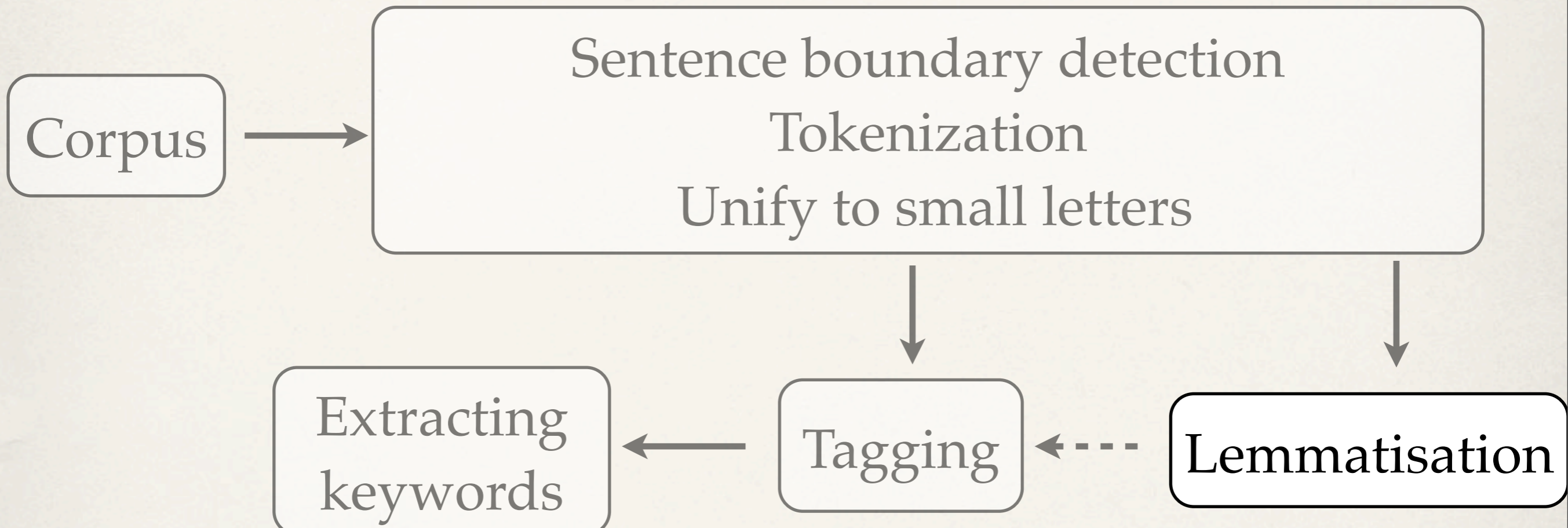
Preprocessing



Preprocessing



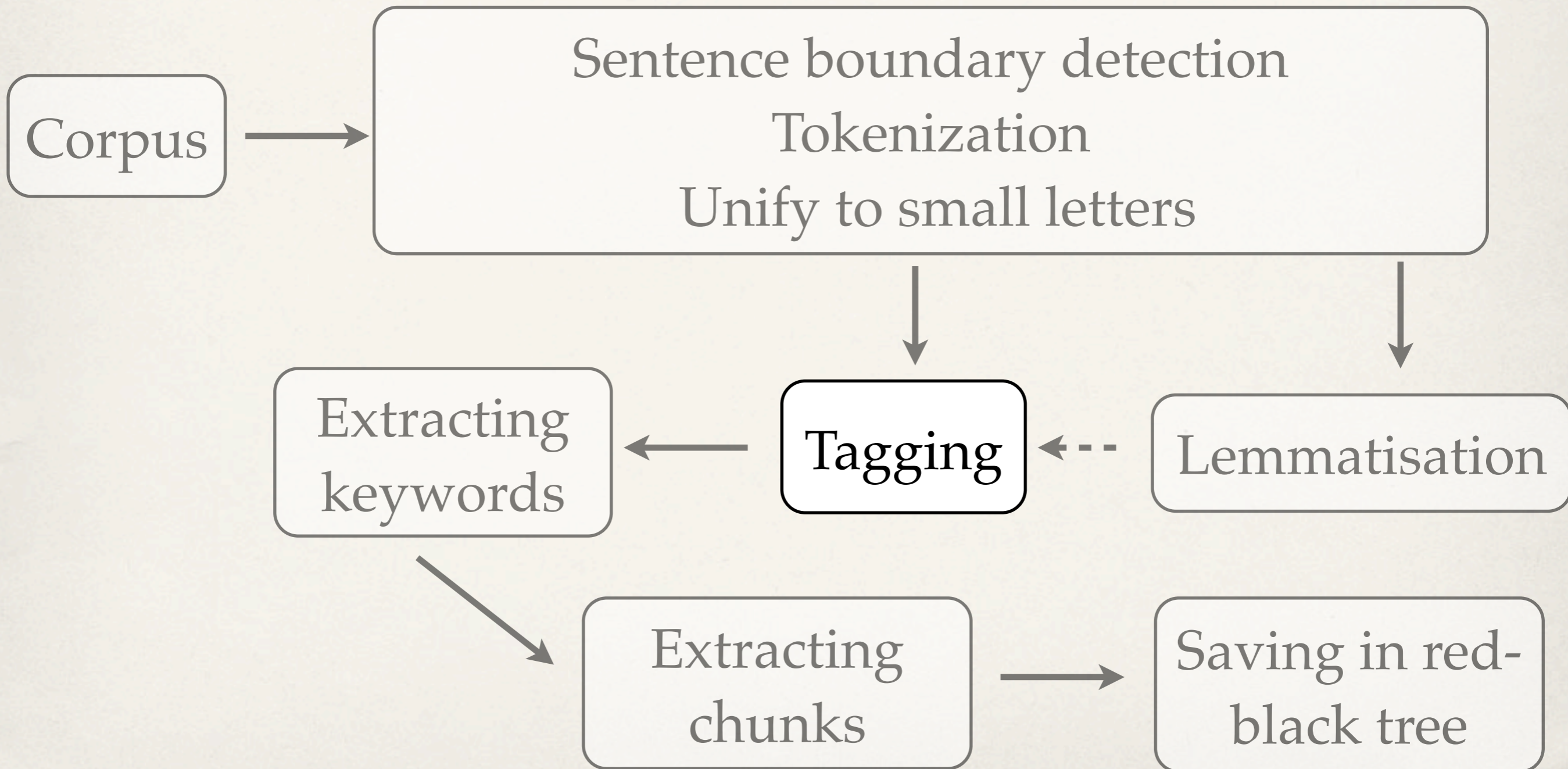
Preprocessing



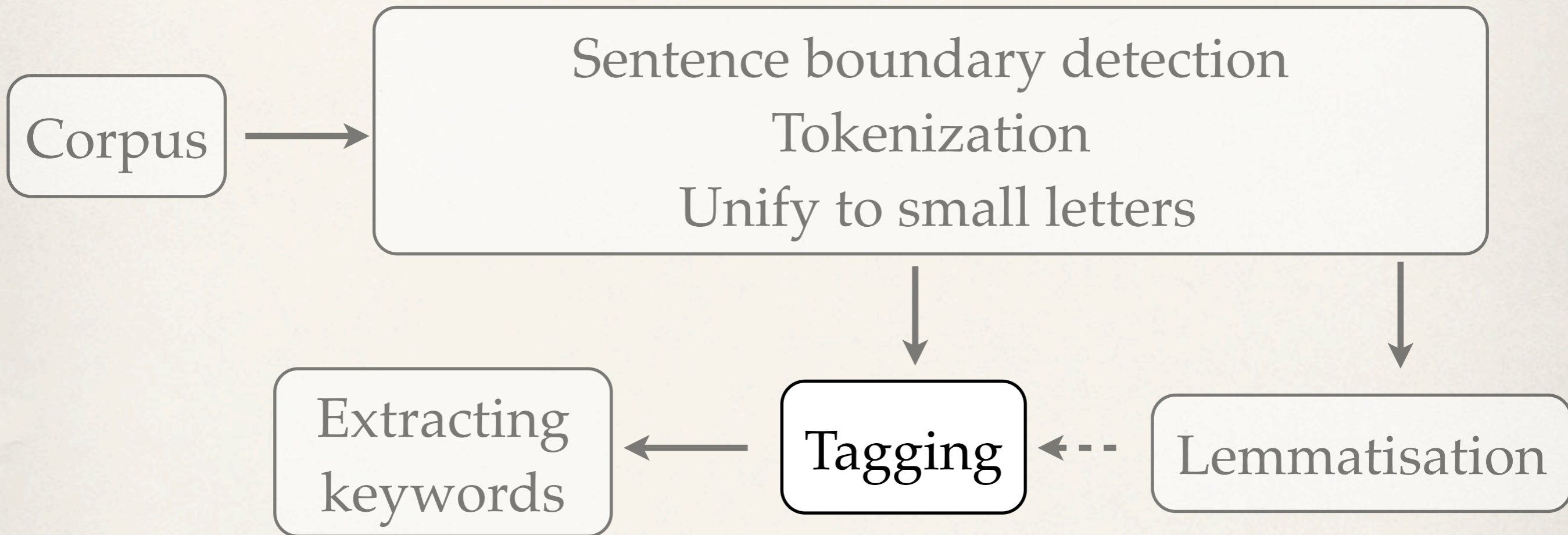
[['the', 'president', 'of', 'cuba', 'invited', 'many',
'**president**', 'to', 'a', 'great', 'dinner', '.'], ['during', 'the',
'dinner', 'they', 'got', 'drunk', 'by', 'wine', ',', 'wine',
'and', 'more', 'wine', '.']]

Saving in red-
black tree

Preprocessing



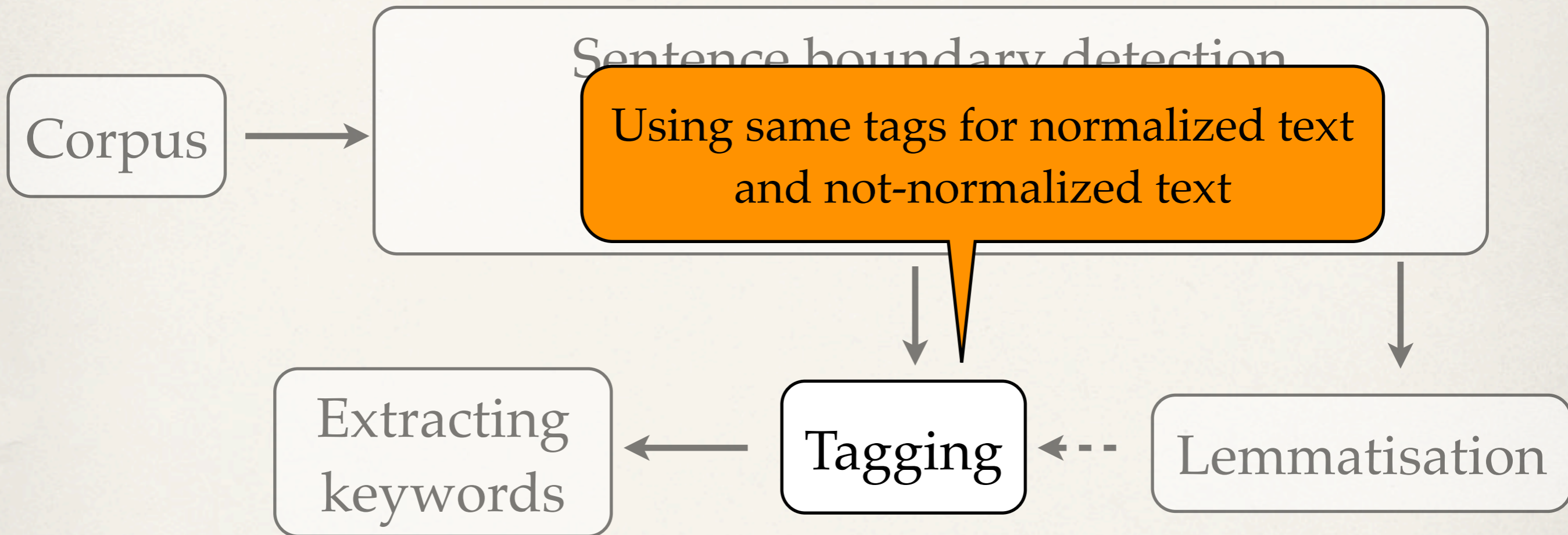
Preprocessing



[[('the', 'DT'), ('president', 'NN'), ('of', 'IN'), ('cuba', 'NN'), ('invited', 'VBN'), ('many', 'JJ'), ('presidents', 'NNS'), ('to', 'TO'), ('a', 'DT'), ('great', 'JJ'), ('dinner', 'NN'), ('.', '.')], [('during', 'IN'), ('the', 'DT'), ...]]

d-

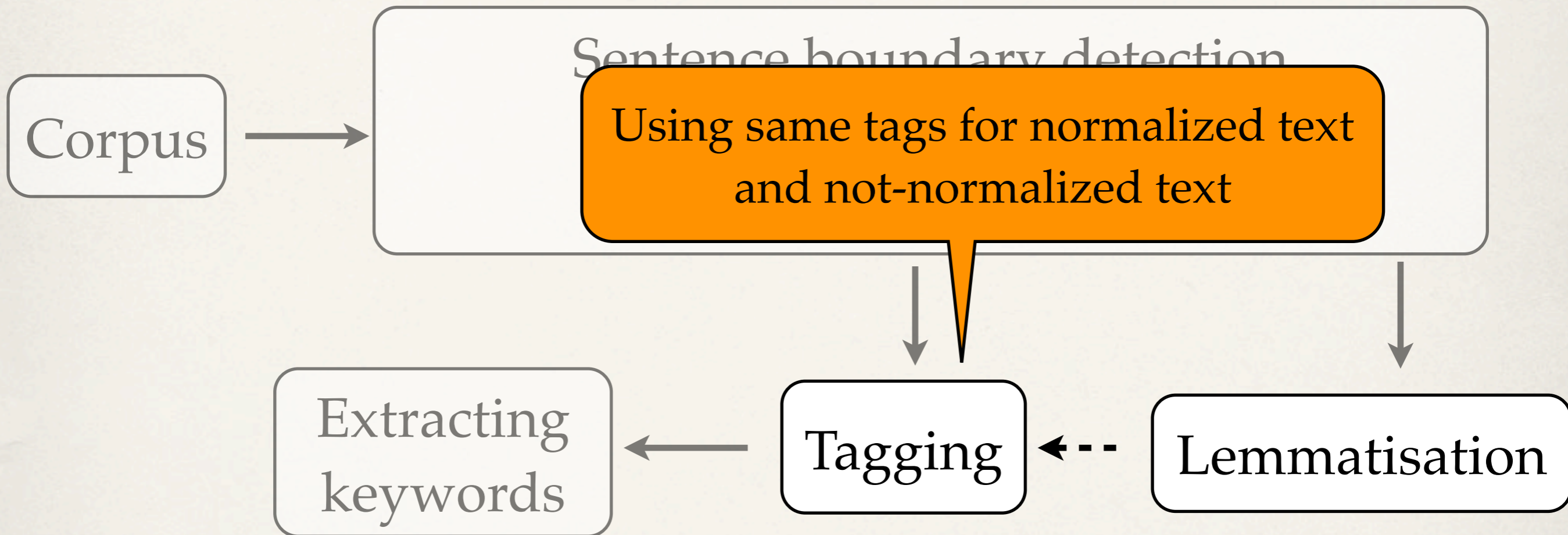
Preprocessing



[[('the', 'DT'), ('president', 'NN'), ('of', 'IN'), ('cuba', 'NN'), ('invited', 'VBN'), ('many', 'JJ'), ('presidents', 'NNS'), ('to', 'TO'), ('a', 'DT'), ('great', 'JJ'), ('dinner', 'NN'), ('.', '.')], [('during', 'IN'), ('the', 'DT'), ...]]

d-

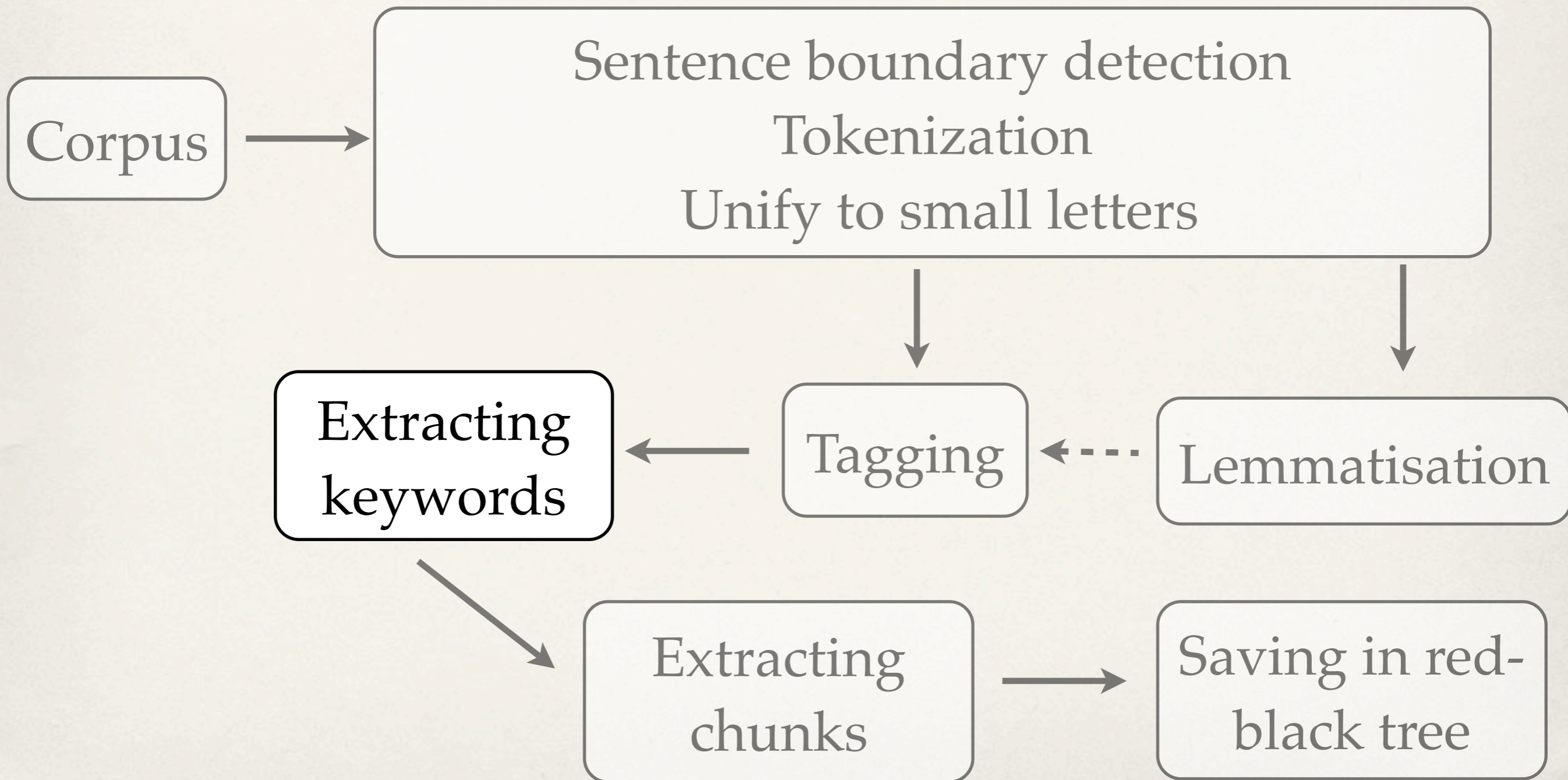
Preprocessing



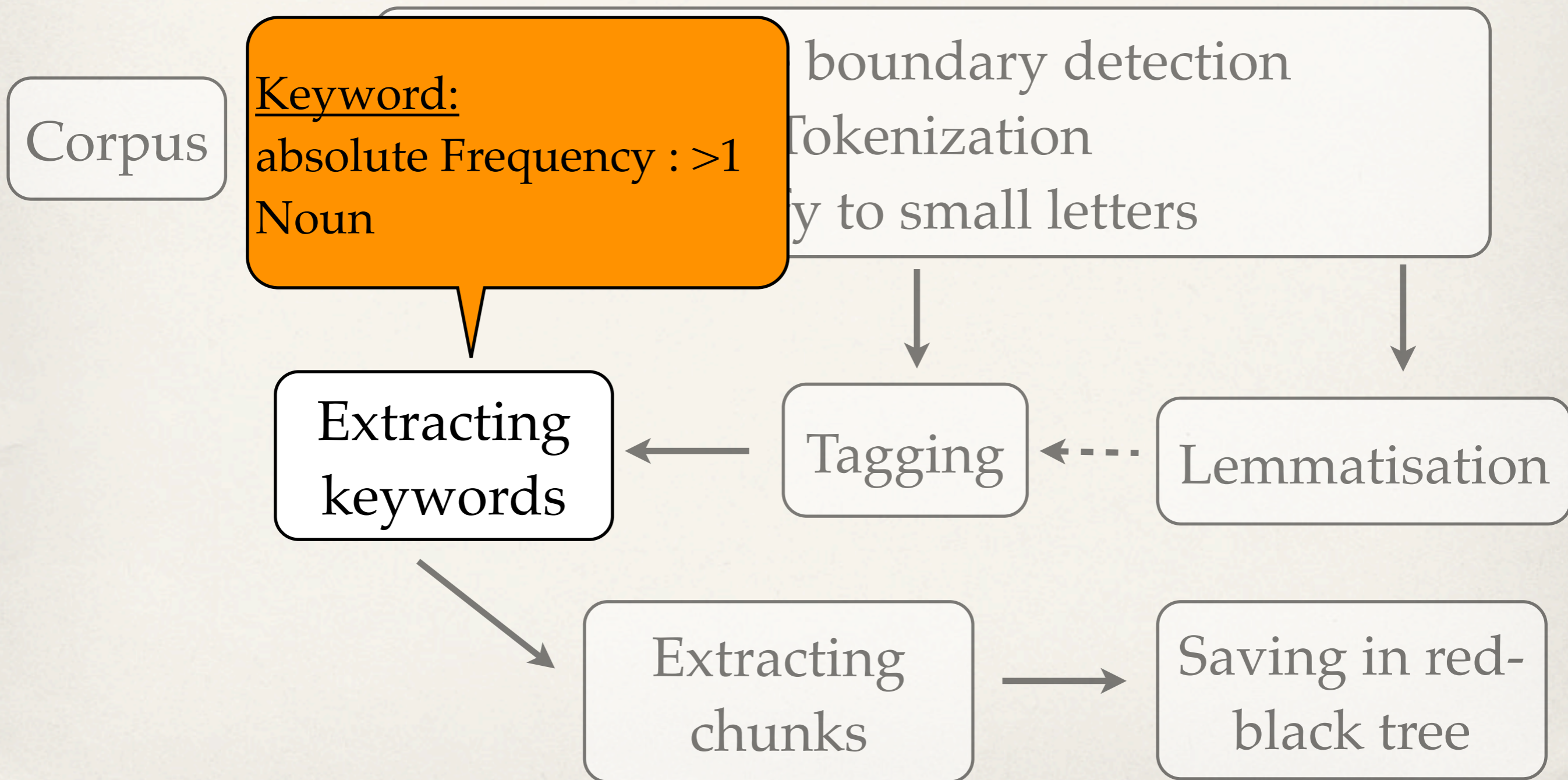
[[('the', 'DT'), ('president', 'NN'), ('of', 'IN'), ('cuba', 'NN'), ('invited', 'VBN'), ('many', 'JJ'), ('**president**', 'NNS'), ('to', 'TO'), ('a', 'DT'), ('great', 'JJ'), ('dinner', 'NN'), ('.', '.')], [('during', 'IN'), ('the', 'DT'), ...]]

d-

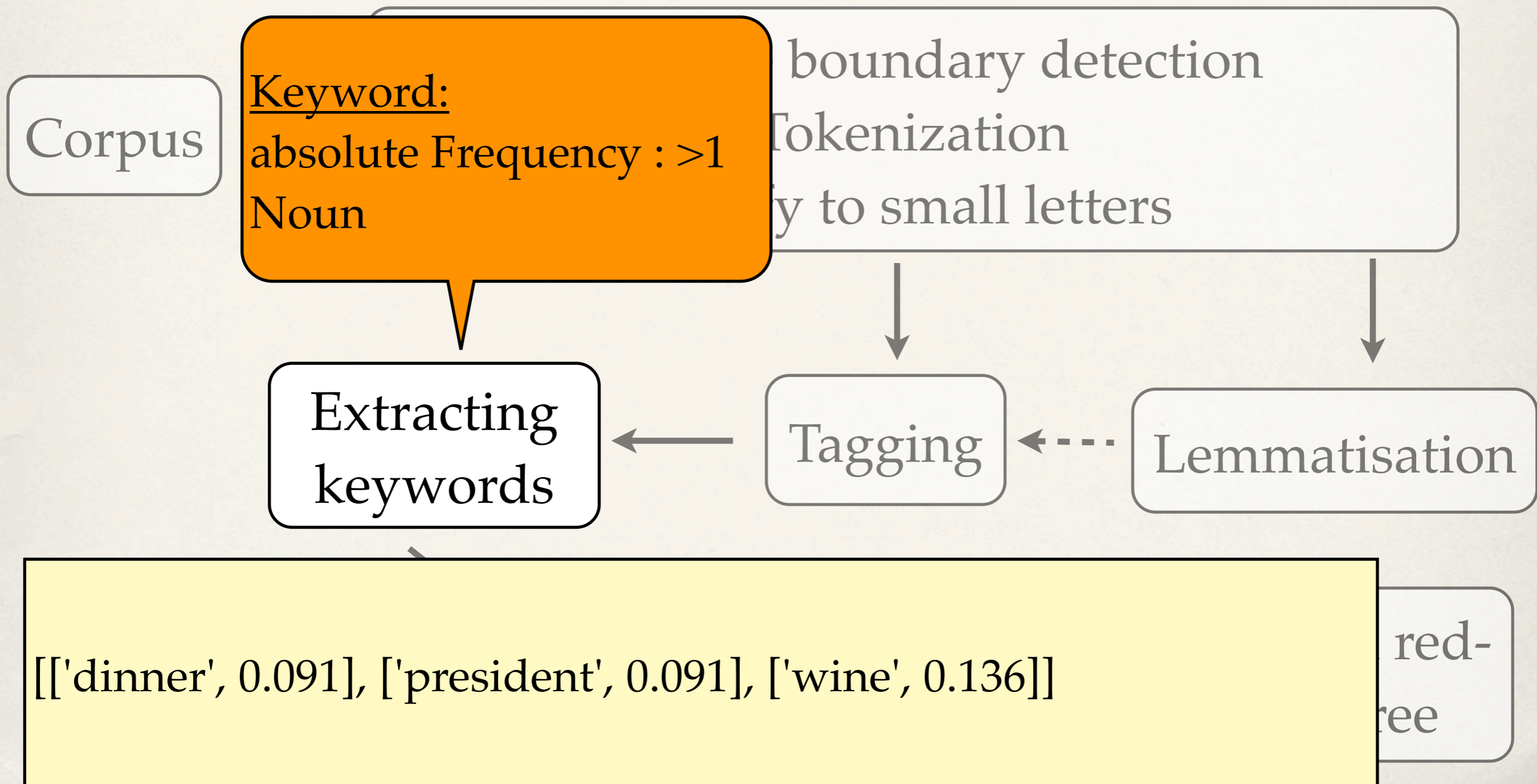
Preprocessing



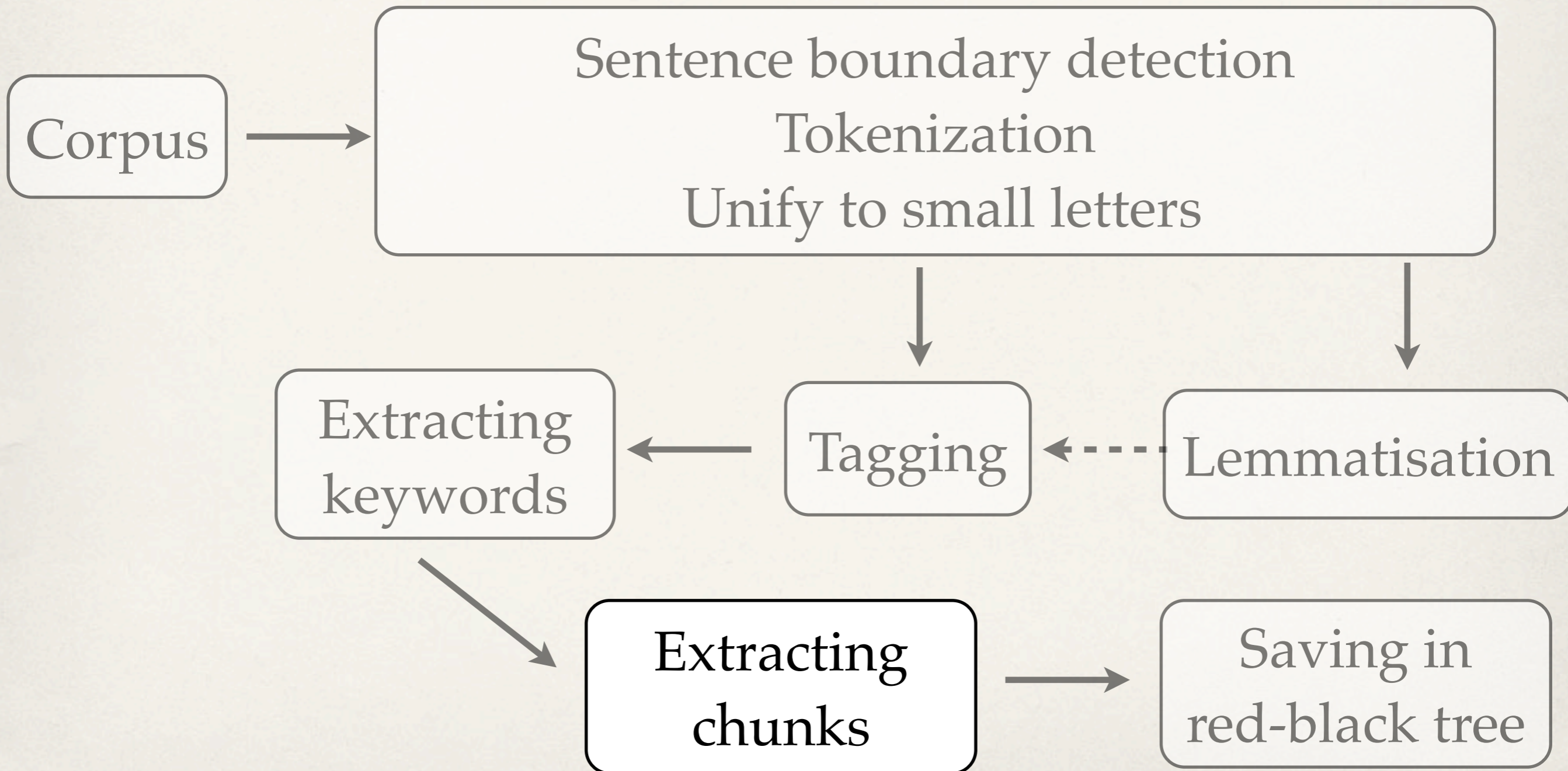
Preprocessing



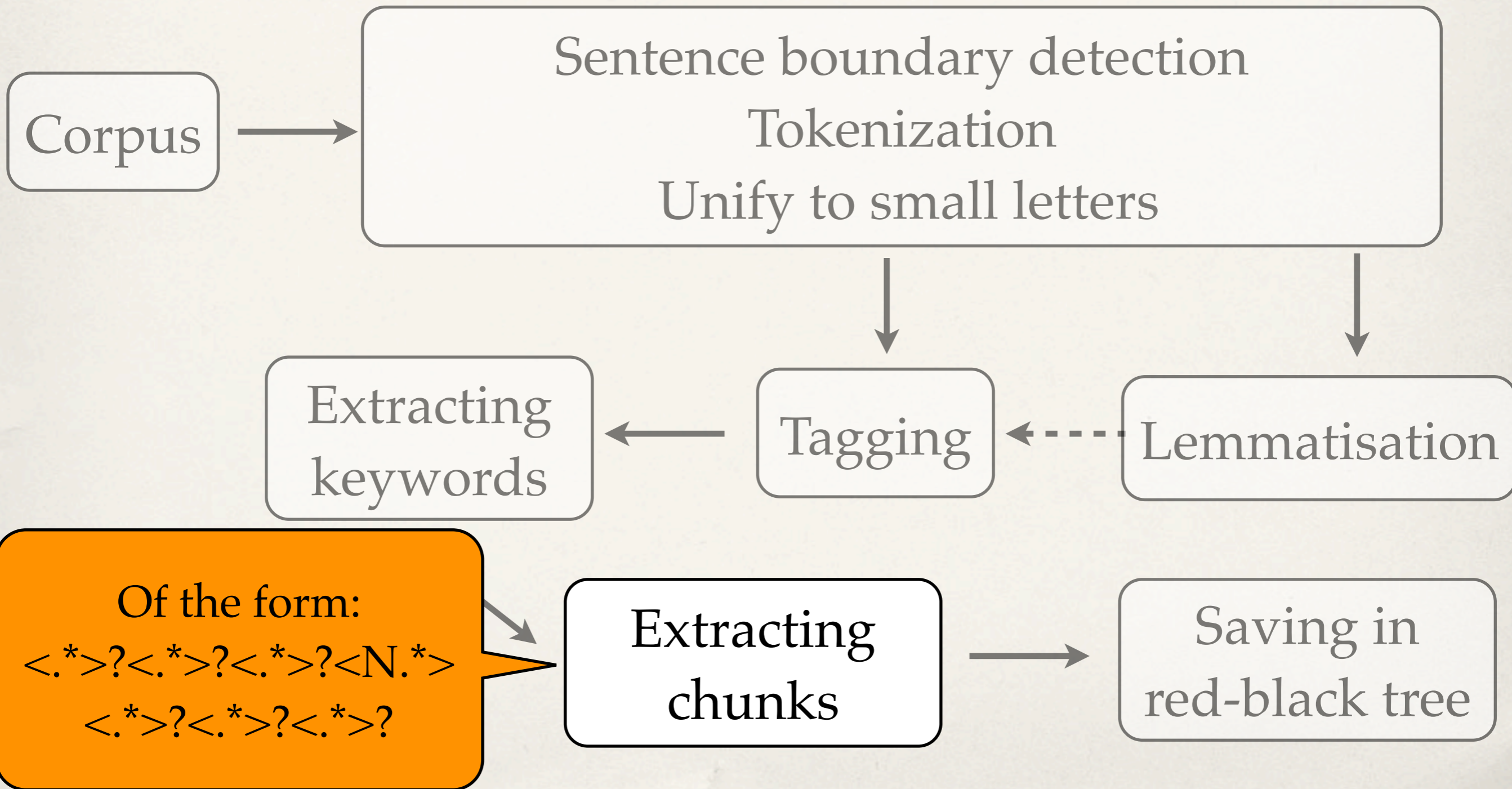
Preprocessing



Preprocessing



Preprocessing



Preprocessing

[during_IN the_DT dinner_NN they_PRP got_VBD drunk_NN]
[to_TO a_DT great_JJ dinner_NN ...]
[the_DT president_NN of_IN cuba_NN invited_VBN]
[cuba_NN invited_VBN many_JJ presidents_NNS to_TO a_DT
great_JJ]
[by_IN wine_NN ,_ wine_NN and_CC more_JJR wine_NN]
[wine_NN and_CC more_JJR wine_NN ...]
[got_VBD drunk_NN by_IN wine_NN ,_ wine_NN and_CC]

keywords

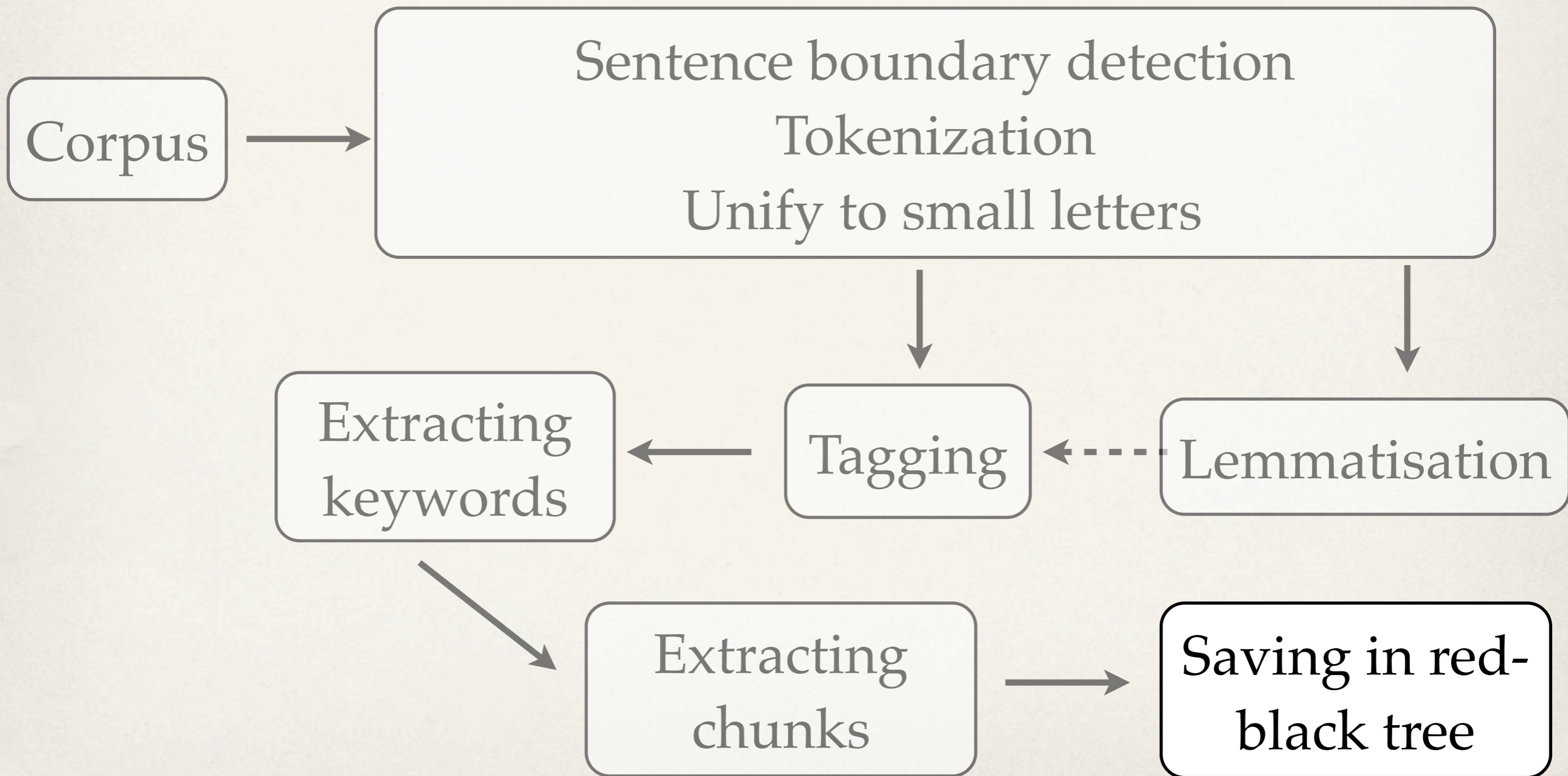
Of the form:

<.*>?<.*>?<.*>?<N.*>
<.*>?<.*>?<.*>?

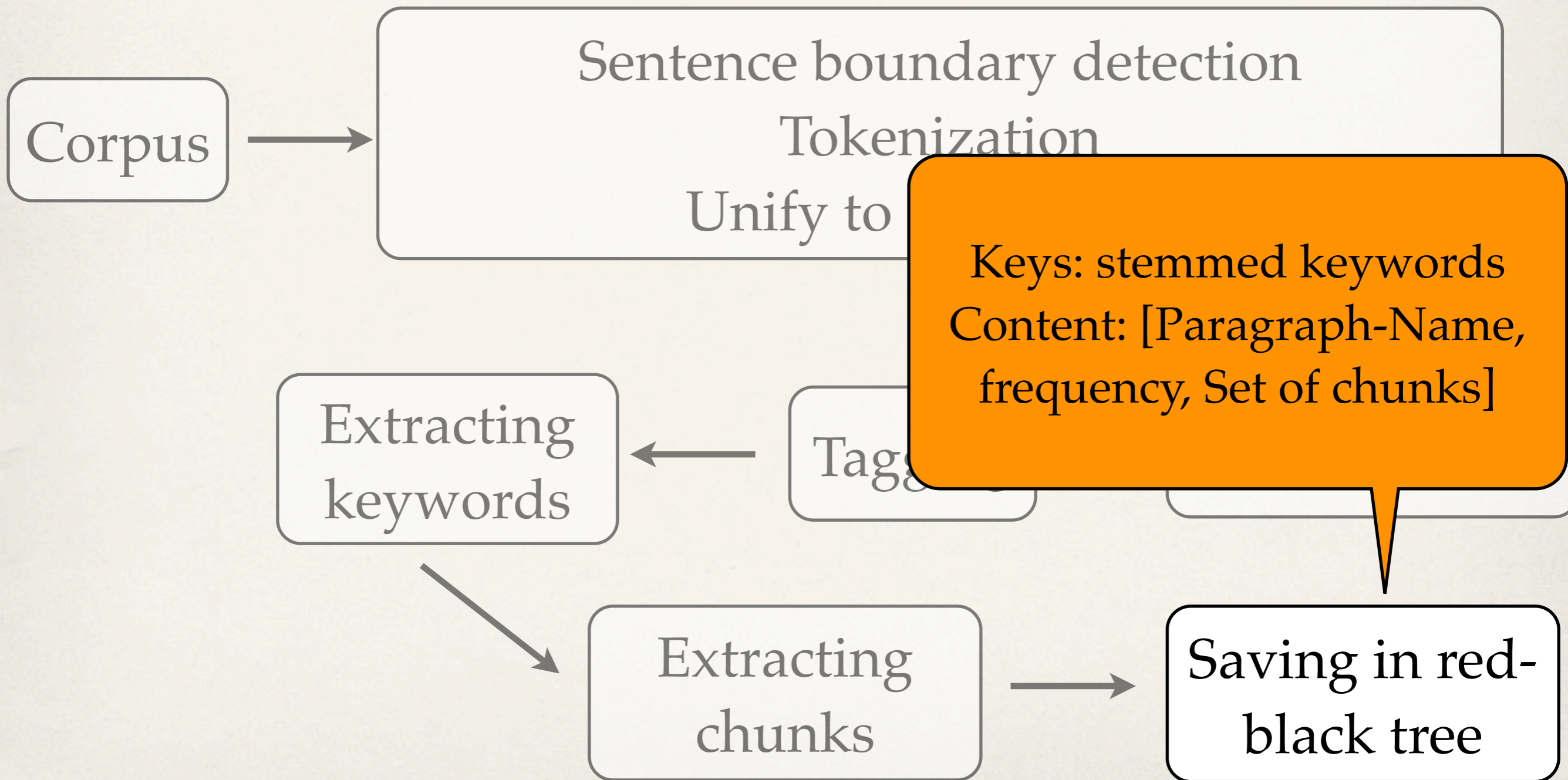
Extracting
chunks

Saving in
red-black tree

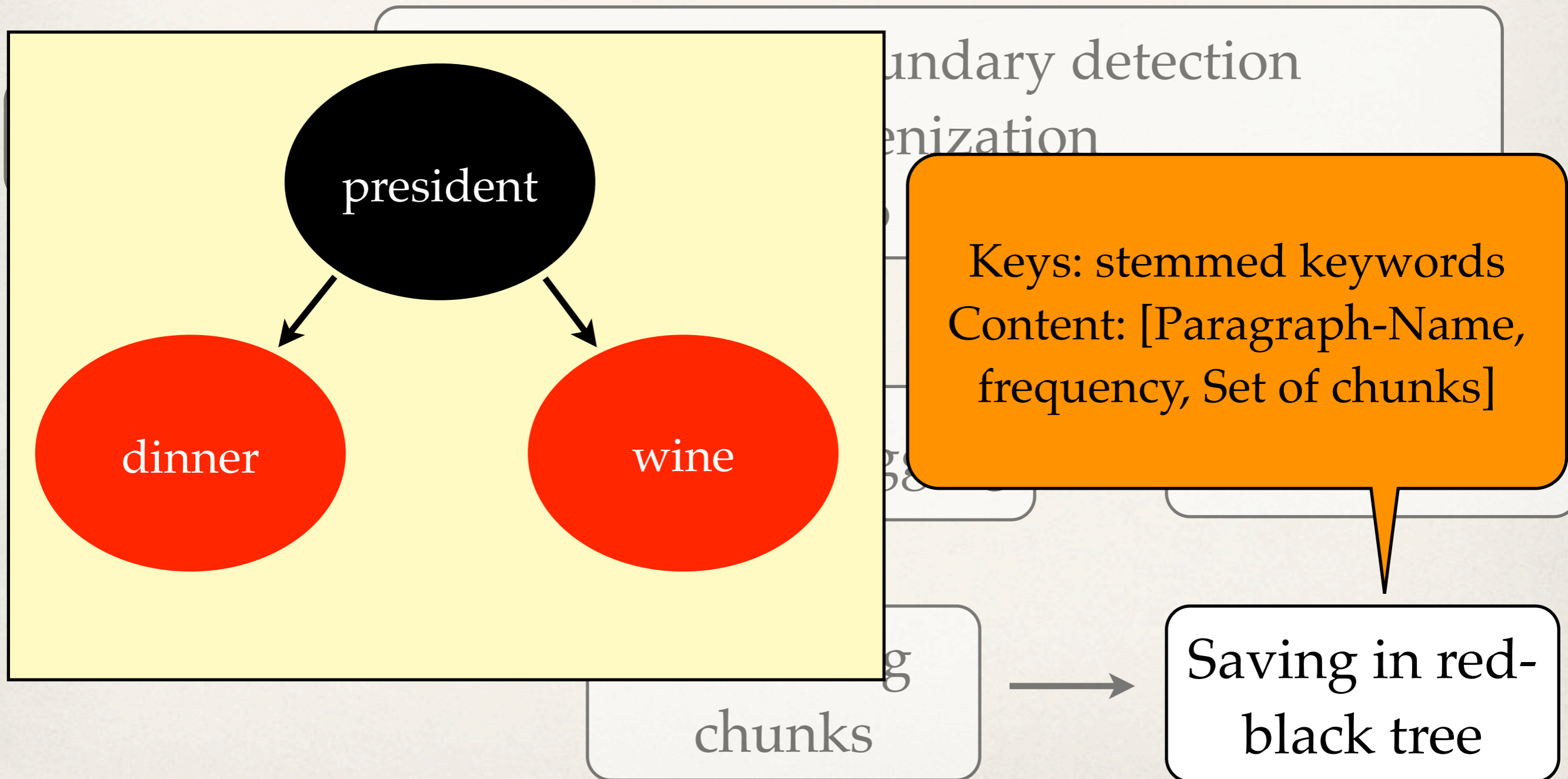
Preprocessing



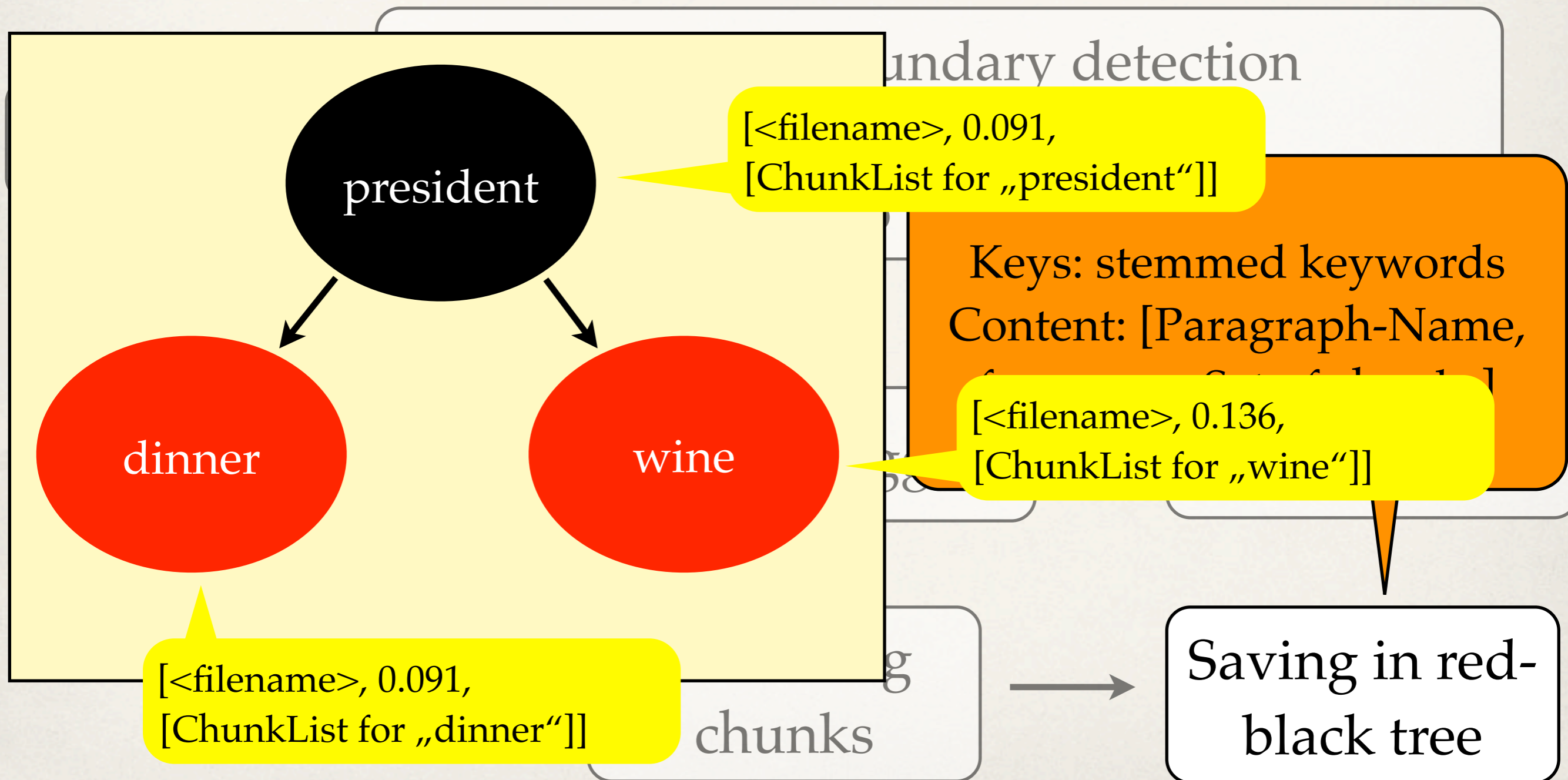
Preprocessing



Preprocessing



Preprocessing



Querying

Fidel Castro's History Summarization

Query
batista

Category
-All Categories-
 General Query
 Load Interviews
 Random Query

Filtered by:
 None
 Y... from: 28.02.201
 Timespan to: 28.02.201

Year Relevance Doc. Details

Selected Docs
Summary

Summary

Chunk 1 : up and said: Batista has fled. Batista h
Chunk 2 : said: Batista has fled. Batista has escap
Then

Document: [19940301_00-062.txt](#)
relative Frequency: 0.018181818

Chunk 1 : to one, in Batista's favor. There was a
Chunk 2 : there were 10,000 Batista soldiers to e

Document: [19940301_00-069.txt](#)
relative Frequency: 0.025

Chunk 1 : rebel army destroyed Batista's offensiv
and raised morale,
Chunk 2 : until we destroyed Batista's offensive.
That is, the

Relevant Paragraphs

DATE:19940301
YEAR:1994
AUTHOR:
HEADLINE:Castro Comments on Radio Rebelde History
PLACE:CARIBBEAN / Cuba
SOURCE:Havana Tele Rebelde and Cuba Vision Networks
FileName:19940301
ID:62
TEXT:63. The balance of forces at its best was 30 soldiers to
**one, in Batista's
favor. There was a moment when there were 10,000 Batista
soldiers to each of**
ours, when we were some six or seven people. Well, that was the
figure in the
beginning [words indistinct] up to 100. The number of our
soldiers increased
vigorously as of May, after the last offensive. After that
offensive came the
Second Front offensive, when many weapons were also seized.
But with such a
correlation of forces, we could not hope to defeat the Army
without the support
of the population. The people had to play their role in the
general strike

Pre-defined Queries
Pre-defined Queries

Timeline for 1959:

Querying

Fidel Castro's History Summarization

Query:

Category: General Query Load Interviews Random Query

Filtered by: None Y... Timespan

from: 28.02.201 to: 28.02.201

Year Relevance Doc. Details

Summary

Chunk 1 : up and said: Batista has fled. Batista h
Chunk 2 : said: Batista has fled. Batista has escap
Then

Document: [19940301_00-062.txt](#)
relative Frequency: 0.018181818

Chunk 1 : to one, in Batista's favor. There was a
Chunk 2 : there were 10,000 Batista soldiers to e

Document: [19940301_00-069.txt](#)
relative Frequency: 0.025

Chunk 1 : rebel army destroyed Batista's offensiv
and raised morale,
Chunk 2 : until we destroyed Batista's offensive.
That is, the

Relevant Paragraphs

DATE:19940301
YEAR:1994
AUTHOR:
HEADLINE:Castro Comments on Radio Rebelde History
PLACE:CARIBBEAN / Cuba
SOURCE:Havana Tele Rebelde and Cuba Vision Networks
FileName:19940301
ID:62
TEXT:63. The balance of forces at its best was 30 soldiers to
**one, in Batista's
favor. There was a moment when there were 10,000 Batista
soldiers to each of**
ours, when we were some six or seven people. Well, that was the
figure in the
beginning [words indistinct] up to 100. The number of our
soldiers increased
vigorously as of May, after the last offensive. After that
offensive came the
Second Front offensive, when many weapons were also seized.
But with such a
correlation of forces, we could not hope to defeat the Army
without the support
of the population. The people had to play their role in the
general strike

Pre-defined Queries

Pre-defined Queries

Timeline for 1959:

Querying

Fidel Castro's History Summarization

Query:

Category: Selected Docs Summary

General Query
 Load Interviews
 Random Query

Filtered by: None Y... from: 28.02.201 to: 28.02.201

Year Relevance Doc. Details

Summary

Chunk 1 : up and said: Batista has fled. Batista h
Chunk 2 : said Batista has fled. Batista has escap
Then

Document: [19940301_00-062.txt](#)
relative Frequency: 0.018181818

Chunk 1 : to one, in Batista's favor. There was a
Chunk 2 : there were 10,000 Batista soldiers to e

Document: [19940301_00-069.txt](#)
relative Frequency: 0.025

Chunk 1 : rebel army destroyed Batista's offensiv
and raised morale,
Chunk 2 : until we destroyed Batista's offensive.
That is, the

Relevant Paragraphs

DATE:19940301
YEAR:1994
AUTHOR:
HEADLINE:Castro Comments on Radio Rebelde History
PLACE:CARIBBEAN / Cuba
SOURCE:Havana Tele Rebelde and Cuba Vision Networks
FileName:19940301
ID:62
TEXT:63. The balance of forces at its best was 30 soldiers to
one, in Batista's
favor. There was a moment when there were 10,000 Batista
soldiers to each of
ours, when we were some six or seven people. Well, that was the
figure in the
beginning [words indistinct] up to 100. The number of our
soldiers increased
vigorously as of May, after the last offensive. After that
offensive came the
Second Front offensive, when many weapons were also seized.
But with such a
correlation of forces, we could not hope to defeat the Army
without the support
of the population. The people had to play their role in the
general strike

Pre-defined Queries

Pre-defined Queries

Timeline for 1959:

Querying

Fidel Castro's History Summarization

Query:

Category: Selected Docs

General Query
 Load Interviews
 Random Query

Filtered by:
 None Y... from: 28.02.201 to: 28.02.201
 Timespan

Year Relevance Doc. Details

Summary

Chunk 1 : up and said: Batista has fled. Batista h
Chunk 2 : said: Batista has fled. Batista has escap
Then

Document: [19940301_00-062.txt](#)
relative Frequency: 0.018181818

Chunk 1 : to one, in Batista's favor. There was a
Chunk 2 : there were 10,000 Batista soldiers to e

Document: [19940301_00-069.txt](#)
relative Frequency: 0.025

Chunk 1 : rebel army destroyed Batista's offensiv
and raised morale,
Chunk 2 : until we destroyed Batista's offensive.
That is, the

Relevant Paragraphs

DATE:19940301
YEAR:1994
AUTHOR:
HEADLINE:Castro Comments on Radio Rebelde History
PLACE:CARIBBEAN / Cuba
SOURCE:Havana Tele Rebelde and Cuba Vision Networks
FileName:19940301
ID:62
TEXT:63. The balance of forces at its best was 30 soldiers to
one, in Batista's
favor. There was a moment when there were 10,000 Batista
soldiers to each of
ours, when we were some six or seven people. Well, that was the
figure in the
beginning [words indistinct] up to 100. The number of our
soldiers increased
vigorously as of May, after the last offensive. After that
offensive came the
Second Front offensive, when many weapons were also seized.
But with such a
correlation of forces, we could not hope to defeat the Army
without the support
of the population. The people had to play their role in the
general strike

Pre-defined Queries

Pre-defined Queries

Timeline for 1959:

Querying

Fidel Castro's History Summarization

Query: batista

Category: -All Categories-
 General Query
 Load Interviews
 Random Query

Filtered by: None
 Y... from: 28.02.201
 Timespan to: 28.02.201

Summary

Chunk 1 : up and said: Batista has fled. Batista h
Chunk 2 : said: Batista has fled. Batista has escap
Then

Document: [19940301_00-062.txt](#)
relative Frequency: 0.018181818

Chunk 1 : to one, in Batista's favor. There was a
Chunk 2 : there were 10,000 Batista soldiers to e

Document: [19940301_00-069.txt](#)
relative Frequency: 0.025

Chunk 1 : rebel army destroyed Batista's offensiv
and raised morale,
Chunk 2 : until we destroyed Batista's offensive.
That is, the

Relevant Paragraphs

DATE:19940301
YEAR:1994
AUTHOR:
HEADLINE:Castro Comments on Radio Rebelde History
PLACE:CARIBBEAN / Cuba
SOURCE:Havana Tele Rebelde and Cuba Vision Networks
FileName:19940301
ID:62
TEXT:63. The balance of forces at its best was 30 soldiers to
**one, in Batista's
favor. There was a moment when there were 10,000 Batista
soldiers to each of**
ours, when we were some six or seven people. Well, that was the
figure in the
beginning [words indistinct] up to 100. The number of our
soldiers increased
vigorously as of May, after the last offensive. After that
offensive came the
Second Front offensive, when many weapons were also seized.
But with such a
correlation of forces, we could not hope to defeat the Army
without the support
of the population. The people had to play their role in the
general strike

Pre-defined Queries

Timeline for 1959:

Querying

Fidel Castro's History Summarization

Query: batista

Category: -All Categories-
 General Query
 Load Interviews
 Random Query

Filtered by:
 None
 Y... from: 28.02.201 to: 28.02.201
 Timespan

Year Relevance Doc. Details

Selected Docs
Summary

Summary

Chunk 1 : up and said: Batista has fled. Batista h
Chunk 2 : said: Batista has fled. Batista has escap
Then

Document: [19940301_00-062.txt](#)
relative Frequency: 0.018181818

Chunk 1 : to one, in Batista's favor. There was a
Chunk 2 : there were 10,000 Batista soldiers to e

Document: [19940301_00-069.txt](#)
relative Frequency: 0.025

Chunk 1 : rebel army destroyed Batista's offensiv
and raised morale,
Chunk 2 : until we destroyed Batista's offensive.
That is, the

Relevant Paragraphs

DATE:19940301
YEAR:1994
AUTHOR:
HEADLINE:Castro Comments on Radio Rebelde History
PLACE:CARIBBEAN / Cuba
SOURCE:Havana Tele Rebelde and Cuba Vision Networks
FileName:19940301
ID:62
TEXT:63. The balance of forces at its best was 30 soldiers to
**one, in Batista's
favor. There was a moment when there were 10,000 Batista
soldiers to each of**
ours, when we were some six or seven people. Well, that was the
figure in the
beginning [words indistinct] up to 100. The number of our
soldiers increased
vigorously as of May, after the last offensive. After that
offensive came the
Second Front offensive, when many weapons were also seized.
But with such a
correlation of forces, we could not hope to defeat the Army
without the support
of the population. The people had to play their role in the
general strike

Pre-defined Queries
 Pre-defined Queries

Timeline for 1959:

Querying

The screenshot shows a web application titled "Fidel Castro's History Summarization" with a Cuban flag header. The interface includes a search bar with the query "batista", a category dropdown set to "Selected Docs" (circled in red), and a "Filtered by:" section with radio buttons for "None", "Y...", and "Timespan". Below the search bar are radio buttons for "Year", "Relevance", and "Doc. Details". The main content area is divided into three sections: "Summary" (showing document chunks and frequencies), "Relevant Paragraphs" (showing a detailed text snippet with red highlights), and "Pre-defined Queries" (with a "Pre-defined Queries" button). A "Timeline for 1959:" slider is visible at the bottom.

Fidel Castro's History Summarization

Query: batista

Category: -All Categories- Selected Docs

Filtered by: None Y... Timespan

Year Relevance Doc. Details

General Query Load Interviews Random Query

Summary

Chunk 1 : up and said: Batista has fled. Batista h
Chunk 2 : said: Batista has fled. Batista has escap
Then

Document: [19940301_00-062.txt](#)
relative Frequency: 0.018181818

Chunk 1 : to one, in Batista's favor. There was a
Chunk 2 : there were 10,000 Batista soldiers to e

Document: [19940301_00-069.txt](#)
relative Frequency: 0.025

Chunk 1 : rebel army destroyed Batista's offensiv
and raised morale,
Chunk 2 : until we destroyed Batista's offensive.
That is, the

Relevant Paragraphs

DATE:19940301
YEAR:1994
AUTHOR:
HEADLINE:Castro Comments on Radio Rebelde History
PLACE:CARIBBEAN / Cuba
SOURCE:Havana Tele Rebelde and Cuba Vision Networks
FileName:19940301
ID:62
TEXT:63. The balance of forces at its best was 30 soldiers to
**one, in Batista's
favor. There was a moment when there were 10,000 Batista
soldiers to each of**
ours, when we were some six or seven people. Well, that was the
figure in the
beginning [words indistinct] up to 100. The number of our
soldiers increased
vigorously as of May, after the last offensive. After that
offensive came the
Second Front offensive, when many weapons were also seized.
But with such a
correlation of forces, we could not hope to defeat the Army
without the support
of the population. The people had to play their role in the
general strike

Pre-defined Queries

Pre-defined Queries

Timeline for 1959:

Discussion

Problems

- ❖ Not precise results on categories like “industry“ due to mismatches on thesaurus data
- ❖ Emphasis on extraction of only statistical data information-domain specific
- ❖ Information about communist ideology and external policy needs more preprocessing e.g filter thesaurus data

Problems

- ❖ No consistent format in the speeches
 - ❖ indistinct passages used both [] or ()
 - ❖ 1994-1996 used completely different format
- ❖ Topic categorization
 - ❖ especially in speeches with mixed topics
 - ❖ In ideology & external relation related speeches

Future Work I

- ❖ Extension of project to extract and summarize more topics (ideology, external relations, specific historical events)
- ❖ Use of different methods for capturing more data e.g NER Recognizer
- ❖ Export data in a relational database for faster & more efficient function of the engine
- ❖ Addition of more features in the Querying Tool
- ❖ Timeline connected to summarization tool

Future Work II

- ❖ Extension of summarizing to capture different data sets important for historians
 - ❖ Diaries of historical personalities
 - ❖ Letters
 - ❖ Interviews, Newspaper archives

References

- ❖ Blei.D. (2009) “Topic Models” Computer Science Department, Princeton University
- ❖ Blei D.M, Ng, Jordan “Laten Dirichlet Allocation“, Journal of Machine Learning(2003) 993-995
- ❖ Topic Modelling Tool:
 - ❖ <http://code.google.com/p/topic-modeling-tool>