Text Processing

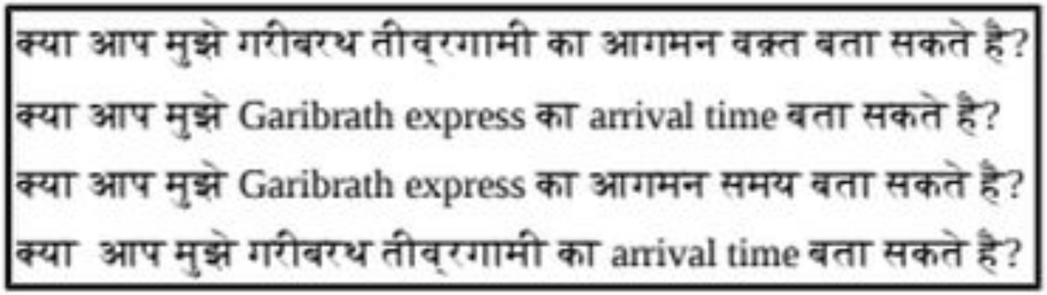
Luciano Barbosa

Centro de Informática
UFPE

Universidade
Federal
de Pernambuco

Influencing Factors

- Who wrote (age, gender, race)
- When was written
- Language: 7097 languages in the world
- Dialects
- Purpose: news, scientific article, book
- Multiple languages: Hindi/English



क्या आप मुझे गरीबरथ तीव्रगामी का आगमन वक़्त बता सकते है?
क्या आप मुझे Garibrath express का arrival time बता सकते है?
क्या आप मुझे Garibrath express का आगमन समय बता सकते है?
क्या आप मुझे गरीबरथ तीव्रगामी का arrival time बता सकते है?

Source:https://kunal-dhawan.weebly.com/language-modelling-of-hindi-english-code-switched-data--a-novel-approach.html

Text Processing Steps

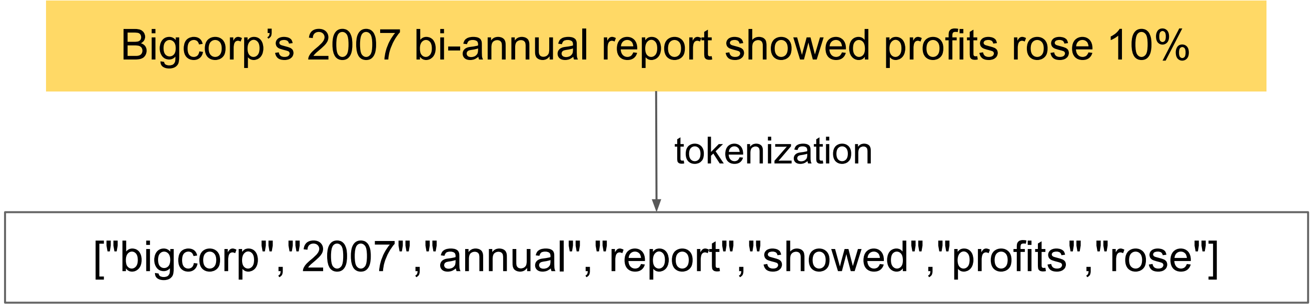
- Tokenization
- Normalization
- Segmentation of sentences

Tokenization

- Brake sentences into words
- How many words are in the spoken sentence?
  - "I do uh main- mainly business data processing"

Tokenization

- Example of simple approach
  - Any sequence of alphanumeric characters of minimum size 3
  - Ended with space or any special character
  - All the letters to lowercase



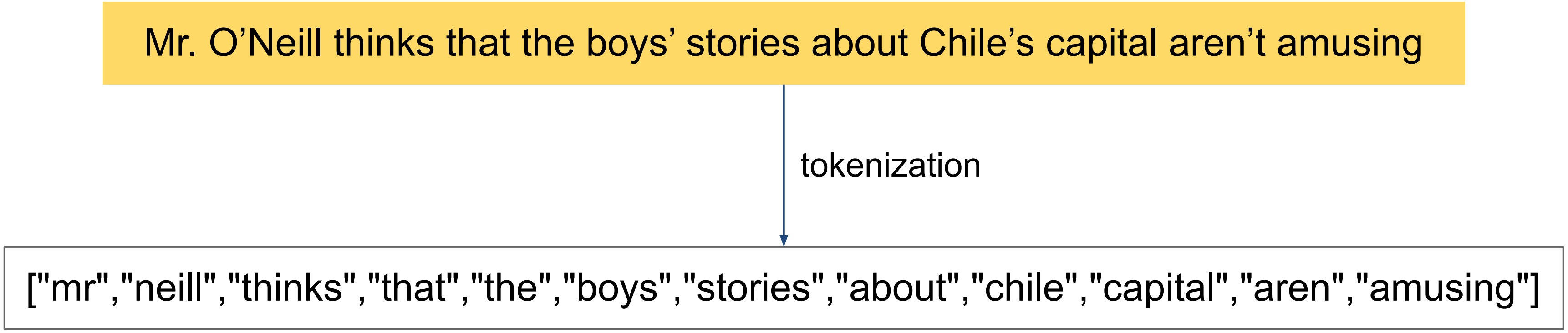
Bigcorp’s 2007 bi-annual report showed profits rose 10%

tokenization

["bigcorp","2007","annual","report","showed","profits","rose"]

Tokenization

- Example of simple approach
  - Any sequence of alphanumeric characters of minimum size 3
  - Ended with space or any special character
  - All the letters to lowercase



Mr. O’Neill thinks that the boys’ stories about Chile’s capital aren’t amusing

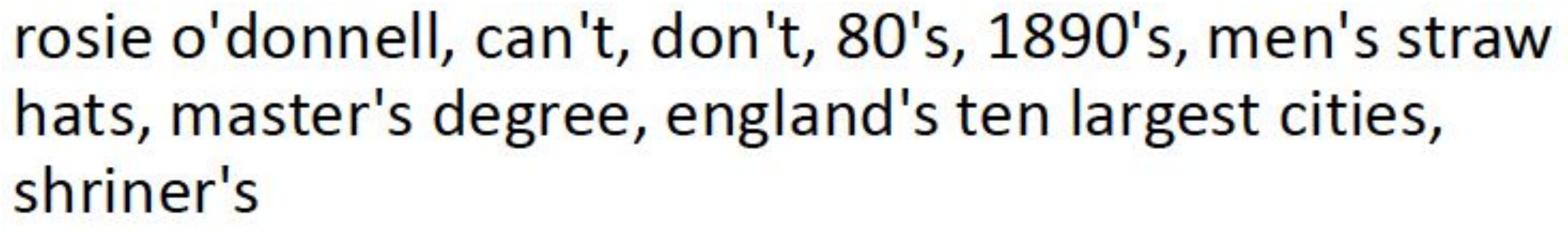
["mr","neill","thinks","that","the","boys","stories","about","chile","capital","aren","amusing"]

Issues in Tokenization

- Small words might be important
  - Ex: am, pm, el (paso), (world war) II
- Hyphens
  - Some times are required
    - Ex: e-bay, wal-mart, cd-rom, t-shirts
  - Separate words
    - Ex: Dallas-Fort Worth, spanish-speaking

Issues in Tokenization

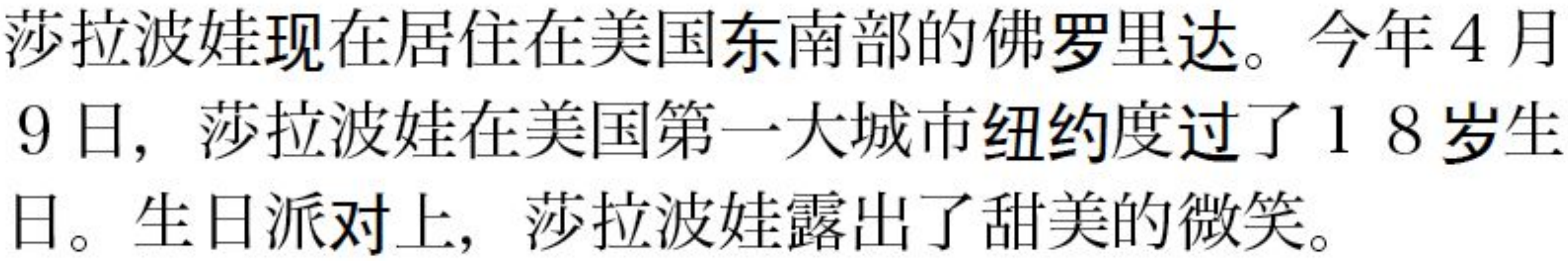
- Special characters are important for URLs, emails, entity names
  - AT&T, $4.5, user@gmail.com
- Word in lowercase can have a different meaning than the original one
  - Bush, Apple
- Apostrophe can be part of a word or part of a possessive (English)


rosie o'donnell, can't, don't, 80's, 1890's, men's straw hats, master's degree, england's ten largest cities, shriner's

Issues in Tokenization

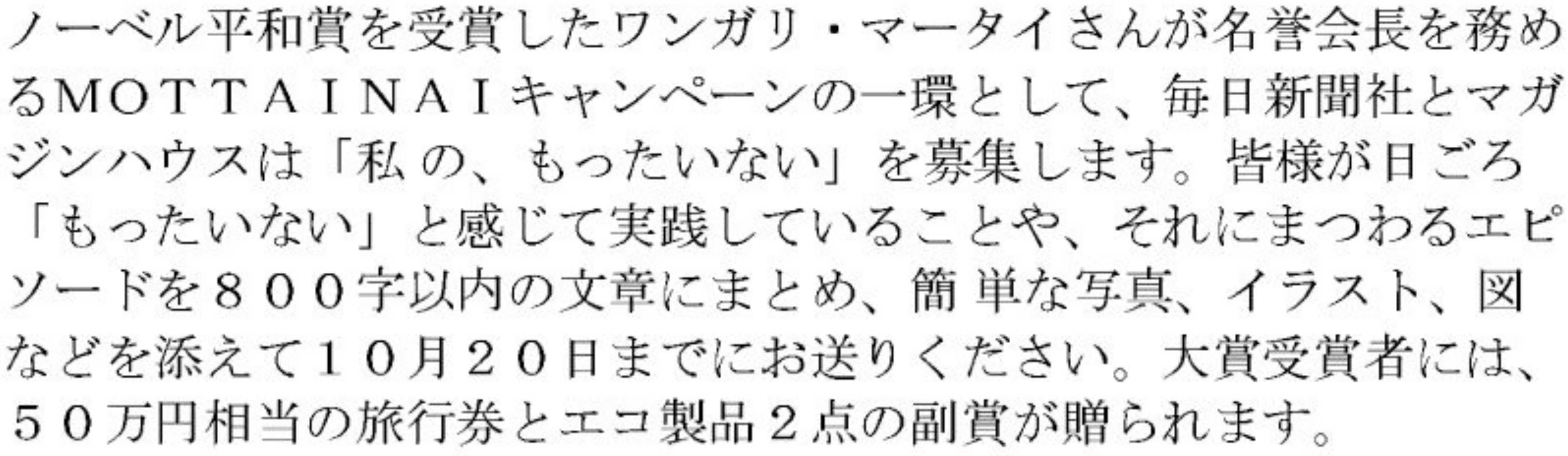
- Number might be useful
  - Ex: nokia 3250, united 93, quicktime 6.5 pro
- Periods can be in numbers, abbreviations, URLs, end of sentences, etc.
  - Ex: I.B.M., Ph.D.

Issues in Tokenization: Chinese



莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

Issues in Tokenization: Japanese



ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるＭＯＴＴＡＩＮＡＩキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを８００字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて１０月２０日までにお送りください。大賞受賞者には、５０万円相当の旅行券とエコ製品２点の副賞が贈られます。

Issues in Tokenization: Chinese

姚明进入总决赛 “Yao Ming reaches the finals”

3 words?

姚明	进入	总决赛
YaoMing	reaches	finals

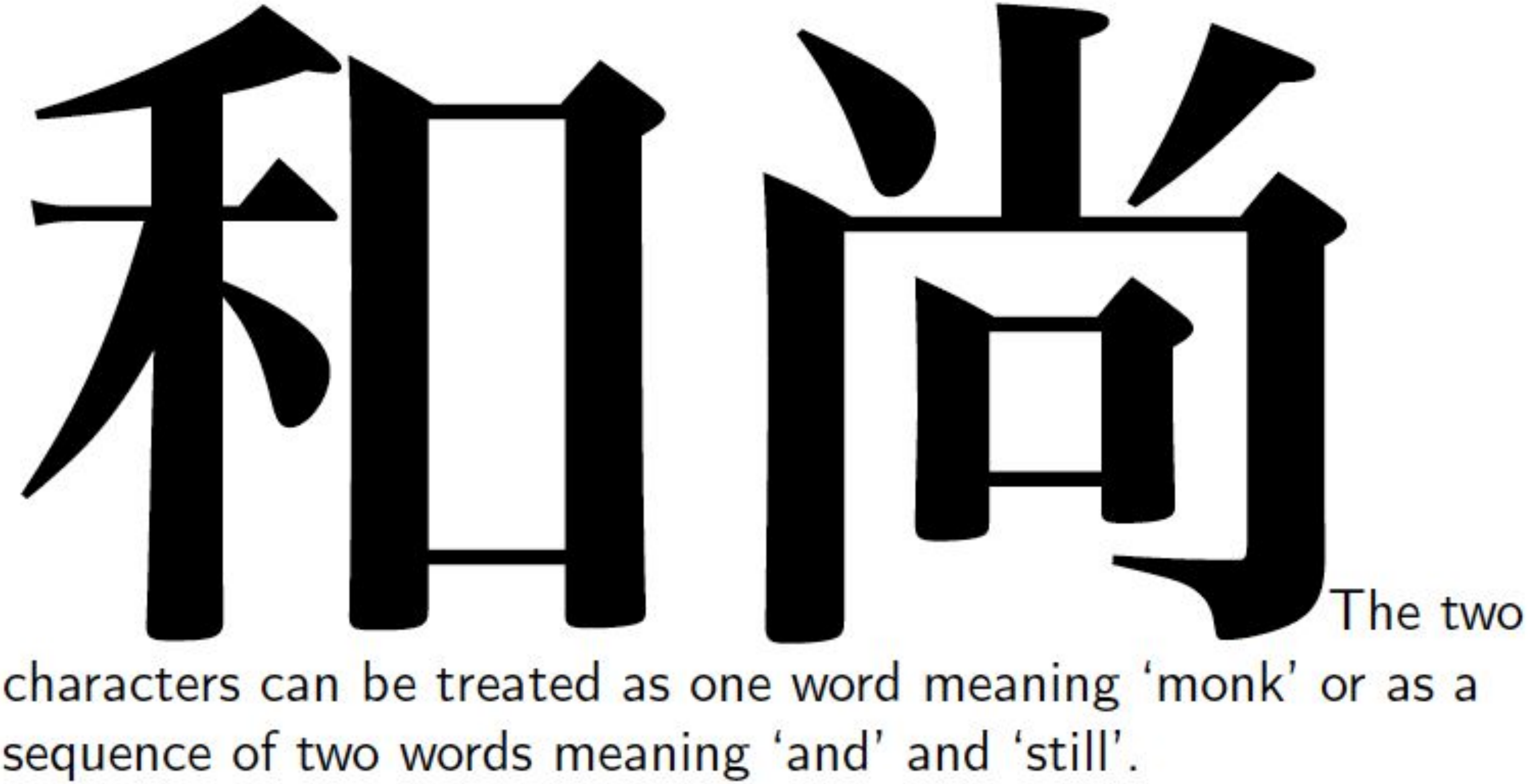
5 words?

姚	明	进入	总	决赛
Yao	Ming	reaches	overall	finals

7 characters?

姚	明	进	入	总	决	赛
Yao	Ming	enter	enter	overall	decision	game

Issues in Tokenization: Ambiguity



和尚 The two characters can be treated as one word meaning ‘monk’ or as a sequence of two words meaning ‘and’ and ‘still’.

Issues in Tokenization: German

computerlinguistik
computer + linguistik
computational linguistics

lebensversicherungsgesellschaftsangestellter
leben + versicherung + gesellschaft + angestellter
life insurance company employee

Issues in Tokenization: Turkish

Uygarlastiramadiklarimizdanmissinizcasina

“(behaving) as if you are among those whom we could not civilize”

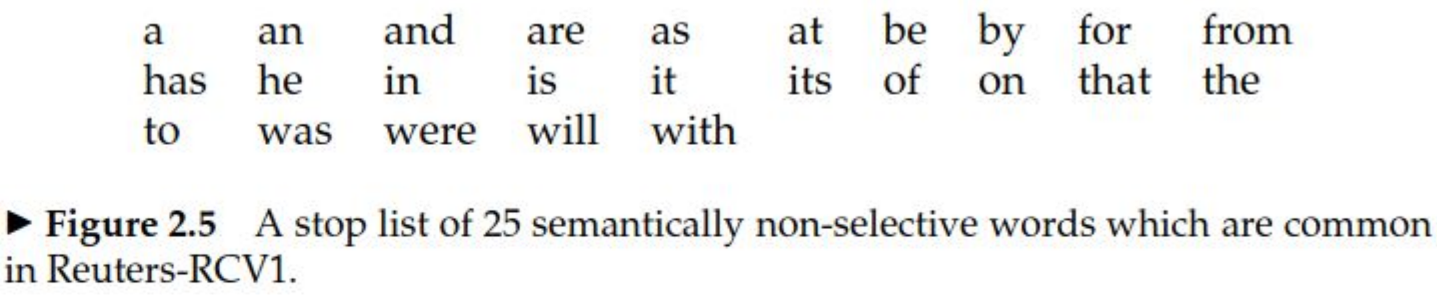
Uygar “civilized” + las “become” + tir “cause” + ama “not able”

+ dik “past” + lar “plural” + imiz “p1pl” +

dan “abl” + mis “past” + siniz “2pl” + casina “as if”

Normalization: Stopwords

- Words with high frequency in a collection
- Don't have much meaning
- Usually not good at differentiating
- Articles, prepositions, conjunctions, etc.
- Created from high frequency words or existing lists



a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

▶ **Figure 2.5** A stop list of 25 semantically non-selective words which are common in Reuters-RCV1.

Fonte: https://nlp.stanford.edu/IR-book/ - Capítulo 2 - The term vocabulary & postings lists

Normalization: Lemmatization

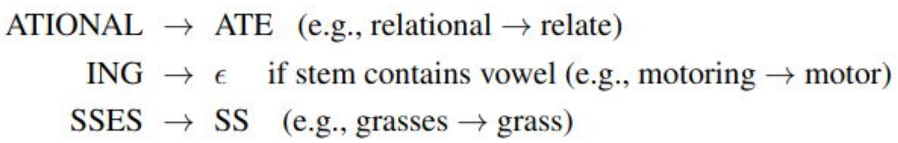
- Group words with the same root (lemma)
  - am, are, is → be
  - car, cars, car's, cars' → car
  - He is reading detective stories → He be read detective story

Normalization: Stemming

- Reduce morphological variations of words to a common stem
- Remove prefix or suffix -> stem
  - Ex: connect – connected, connecting, connection, connections
- There is no consensus on benefits (depends on the language/task)

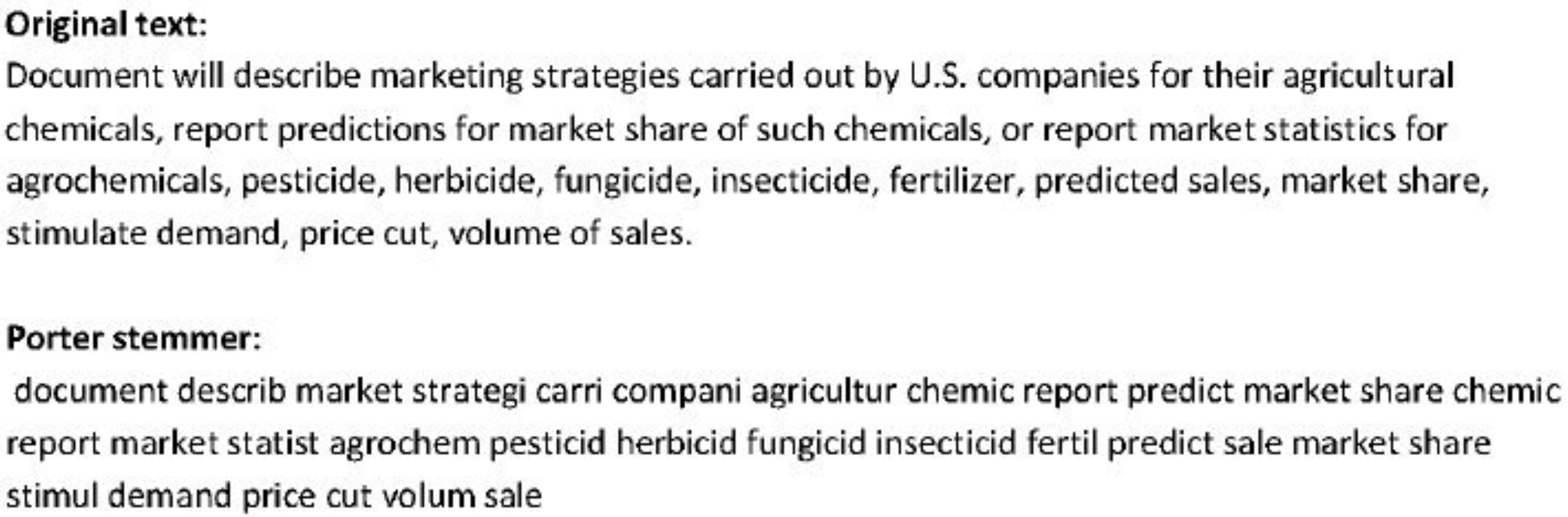
Porter Stemmer

- Very popular
- Consist of a series of rules
- Make mistakes and hard to modify



ATIONAL → ATE (e.g., relational → relate)
ING → ε if stem contains vowel (e.g., motoring → motor)
SSES → SS (e.g., grasses → grass)

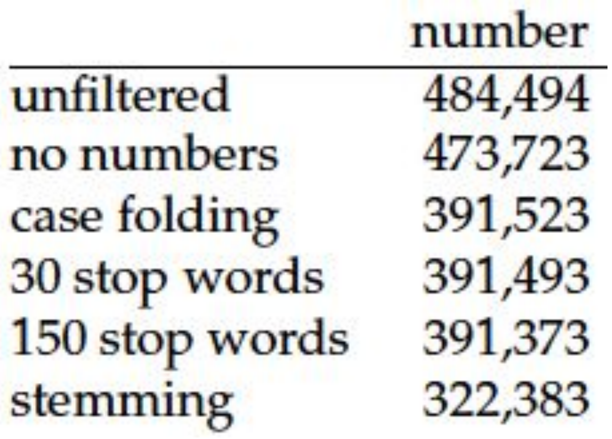
Example of rules



Original text:
Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Porter stemmer:
document describ market strategi carri compani agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale market share stimul demand price cut volum sale

Effect of Normalization on Reuters-RCV1



	number
unfiltered	484,494
no numbers	473,723
case folding	391,523
30 stop words	391,493
150 stop words	391,373
stemming	322,383

Sentence Segmentation

- Characters ! and ? are precise to separate sentences
- Character "." not so much
  - Abbreviation: Dr.
  - Number: 7.5%
- Most used strategy:
  - Tokenize first
  - Uses rules or Machine Learning to classify the point
  - Use of abbreviation dictionary

Tools and Libraries

- Stanford tokenizer para inglês (http://nlp.stanford.edu/software/tokenizer.shtml)
- CoreNLP (https://stanfordnlp.github.io/CoreNLP/)
- NLTK (https://www.nltk.org/)
- Spacy (https://spacy.io/)