OPINION MINING AND ITS IMPORTANCE IN THE ACTIVITY OF CONTEMPORARY ENTERPRISES

Paweł Lula

Cracow University of Economics

Poland

Opinion mining vs. sentiment analysis

- ***Opinion mining:***
  - processing a set of search results for a given item,
  - generating a list of product attributes (quality, features, etc.),
  - aggregating opinion about them (poor, mixed, good).

Source: Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW, pages 519–528, 2003.

Opinion mining vs. sentiment analysis

- ***Sentiment***:
  - overall opinion towards the subject matter.

Source: Thumbs up? Sentiment classification using machine learning techniques, Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, Proceedings of EMNLP, pp. 79--86, 2002

Difficulties in opinion mining

- Hidden character of emotions and sentiment (very often they are not expressed directly),
- Sarcasm (irony),
- Mistakes in writing,
- Repetitions of letters, words and phrases,
- Co-references – two or more expressions refer to the same person or thing (*mobile, phone, headphone, it, ...*),
- Negations (*dislike*, *do not like*),
- Entity recognition problem – identification of names of persons, organizations, locations, monetary values,
- Comparisons (... better then...),
- Polysemy (book = reserve, book = text).

Information retrieval vs. opinion mining

- Information retrieval – identification and analysis ***objective*** pieces of data,
- Opinion mining (sentiment analysis) – identification and analysis ***subjective*** opinions, emotions and feelings.

Opinions’ objectivisation

- ***Voting*** – taking into account a large number of opinions, opinion aggregation,
- ***Opinion evaluation*** – by other customers,
- ***Evaluation of author’s authority*** – publishing some details about opinion’s author and evaluation of his/her authority by others.

Opinions’ taxonomy

- ***Criterion: opinion’s form***
  - ***Binary*** opinions (yes/no, like/dislike, good/bad),
  - ***Nominal*** values (about mobile phone: heavy, expensive, modern, ...),
  - ***Ordered*** values (bad/typical/good/excellent; Likert scale)
  - ***Text***:
    - structured opinions,
    - unstructured opinions.
- ***Criterion: the scope of knowledge***
  - without additional domain knowledge (based only on opinions),
  - with additional domain knowledge (based on opinions and on knowledge about products or services).

Types of analysis (goals of analysis)

- **Sentiment recognition** – analysis the general attitude to a product (*positive, negative, neutral*),
- **Feature-based analysis** – identification and evaluation of main features of a given product.

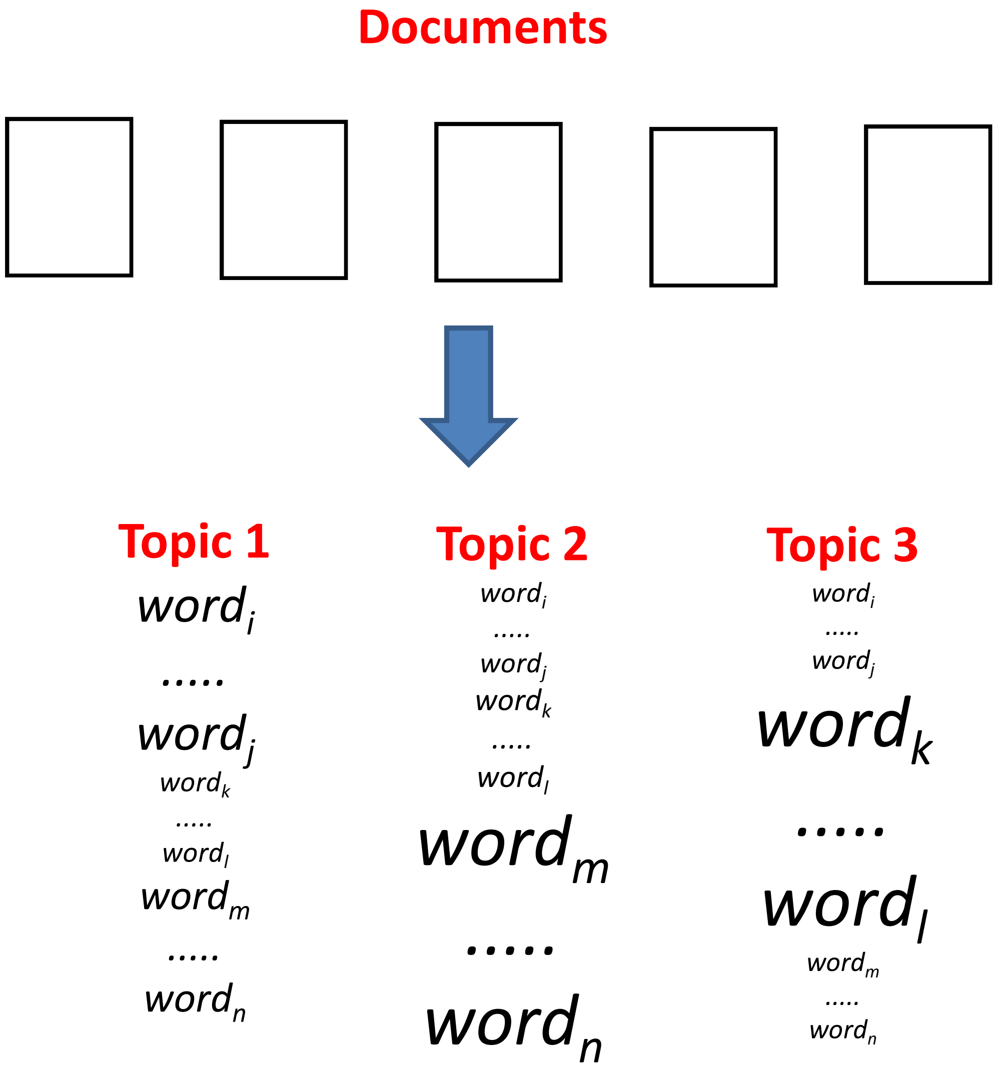
Methods of analysis

- approach based on frequency matrix,
- probabilistic approach (topic modelling, probabilistic LSA),
- rule-based methods (used regular expressions),
- approach based on domain knowledge (ontology-based approach, logic models),
- summarisation and keywords identification methods,
- classification methods – for sentiment classification,
- visualisation methods,
- aggregation methods – used for increasing the level of objectivism.

Examplary solution

- Opinions about hotel rooms in London (Source: http://kavita-ganesan.com/opinosis-opinion-dataset),
- Feature-based analysis,
- Model of domain knowledge: list of attributes,
- Method: probabilistic model – (topic modelling based on Labeled Latent Dirichlet Allocation).

Latent Dirichlet Allocation (LDA)



Documents

Topic 1

word$_i$

.....

word$_j$

word$_k$

.....

word$_l$

word$_m$

.....

word$_n$

Topic 2

word$_i$

.....

word$_j$

word$_k$

.....

word$_l$

word$_m$

.....

word$_n$

Topic 3

word$_i$

.....

word$_j$

word$_k$

.....

word$_l$

word$_m$

.....

word$_n$

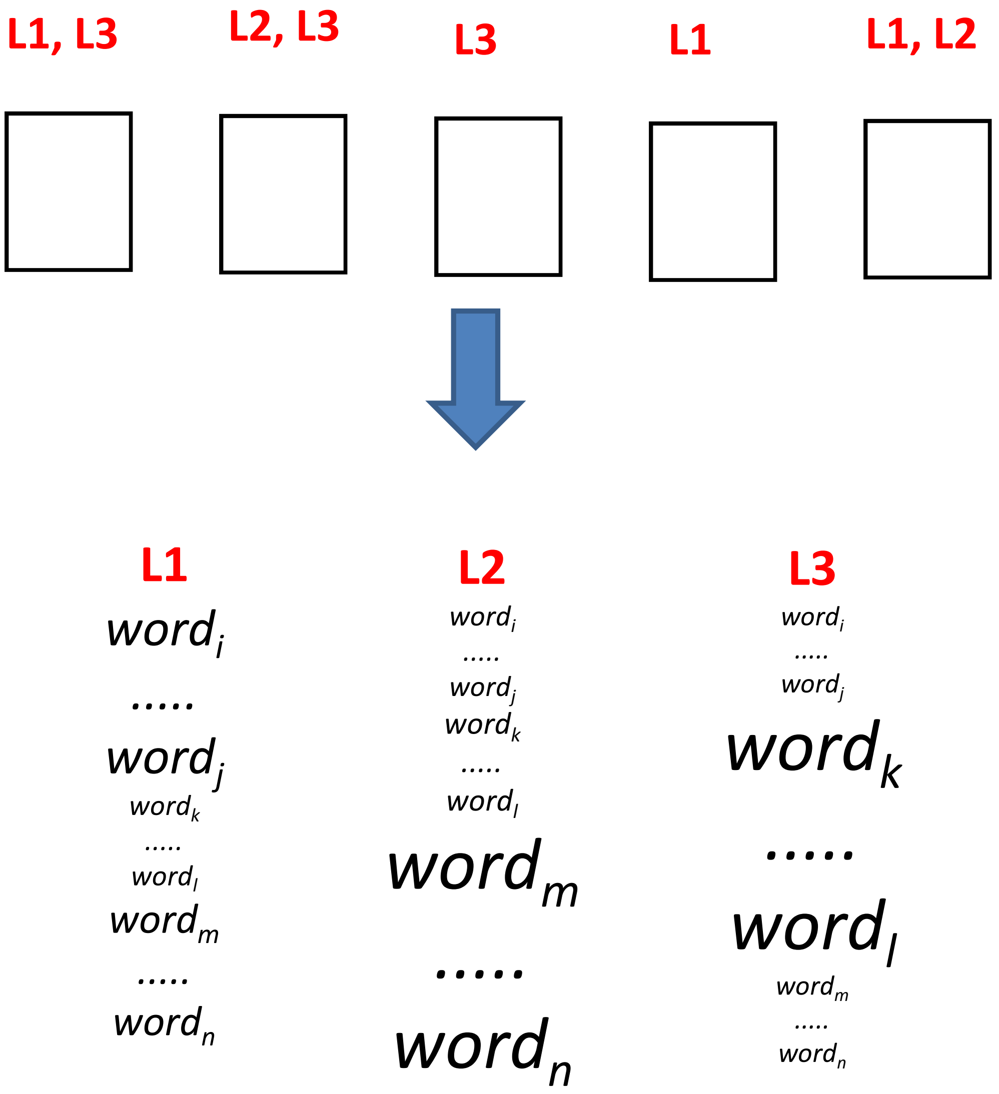
***Latent Dirichlet Allocation** – completely **unsupervised** method of topics identification.*

Topics are described in terms of discrete probabilities over words.

Each document can be modeled as a mixture of topics.

*Topics are **hard** to interpret.*

Labeled-LDA (Ramage, Hall, Nallapati, Manning - 2009)



L1, L3

L2, L3

L3

L1

L1, L2

L1

word$_i$
.....
word$_j$
word$_k$
.....
word$_l$
word$_m$
.....
word$_n$

L2

word$_i$
.....
word$_j$
word$_k$
.....
word$_l$
word$_m$
.....
word$_n$

L3

word$_i$
.....
word$_j$
word$_k$
.....
word$_l$
word$_m$
.....
word$_n$

Labeled Latent Dirichlet Allocation – supervised *method of topic creation.*

Topics ***represent*** *labels (concepts) used for documents' tagging (number of topics = number of different labels).*

Topics are described in terms of discrete probabilities over words.

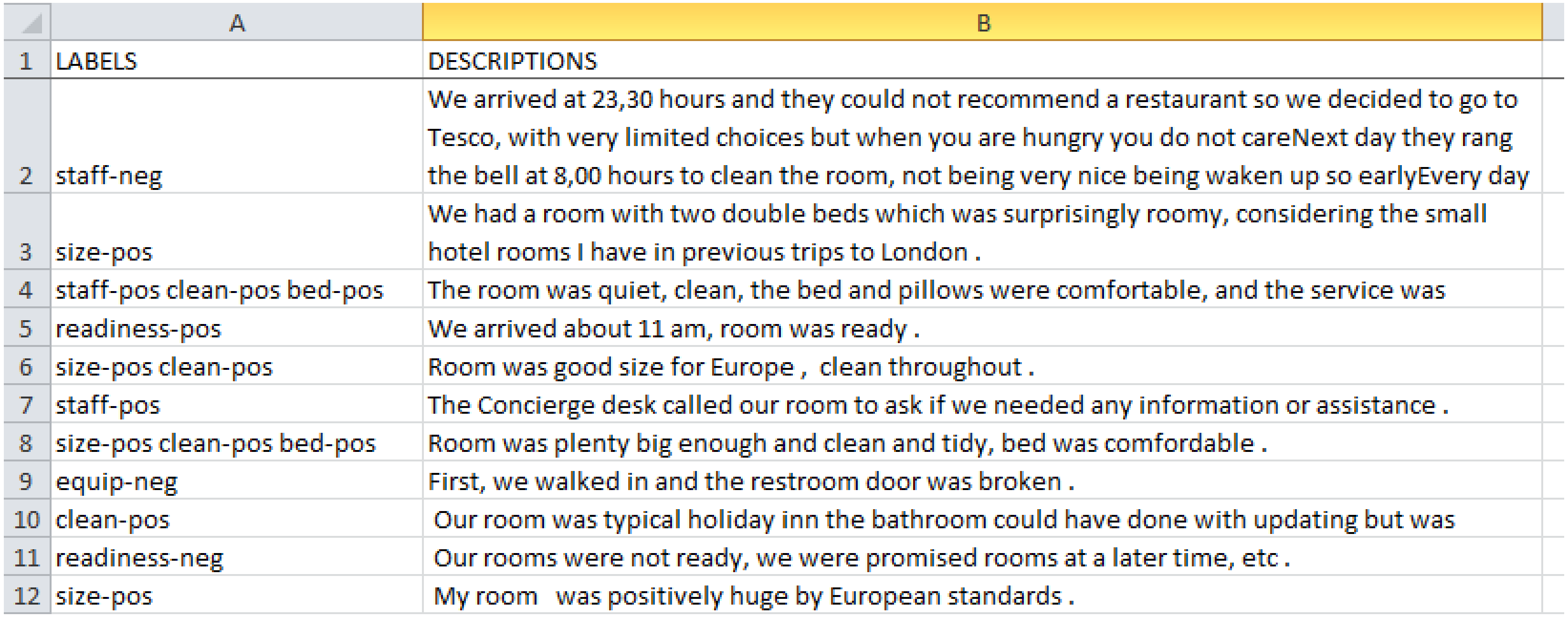
Each document can be modeled as a mixture of topics.

Topics are ***easy*** *to interpret.*

Model of domain knowledge – a set of two-state atributes

- general
- size
- staff
- bed
- clean
- equip
- readiness
- location
- quiet
- comfort
- light
- internet
- price
- temperature
- bathroom
- food
- view
- secure
- decor
- reservation

Data set structure

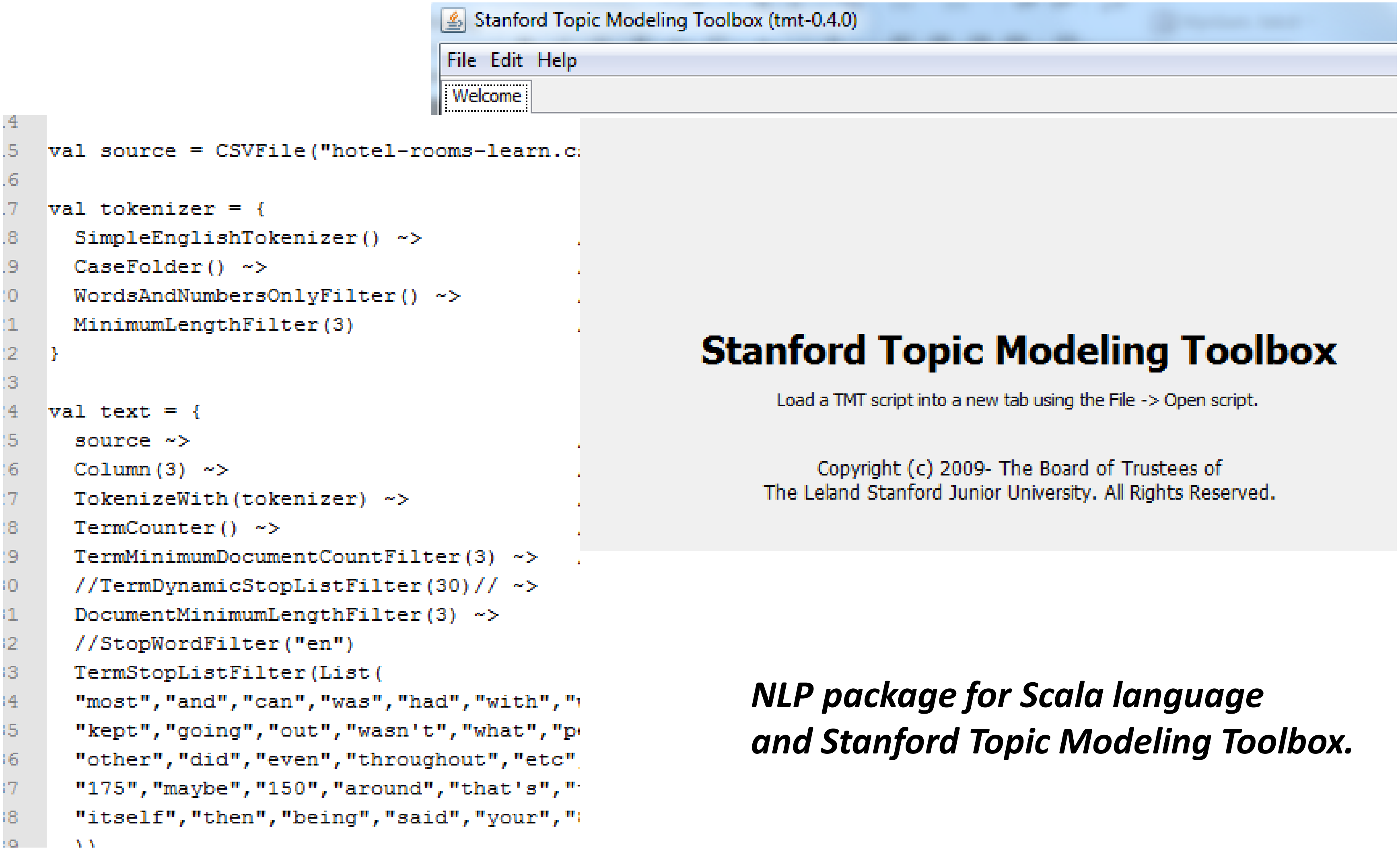


	A	B
1	LABELS	DESCRIPTIONS
2	staff-neg	We arrived at 23,30 hours and they could not recommend a restaurant so we decided to go to Tesco, with very limited choices but when you are hungry you do not careNext day they rang the bell at 8,00 hours to clean the room, not being very nice being waken up so earlyEvery day
3	size-pos	We had a room with two double beds which was surprisingly roomy, considering the small hotel rooms I have in previous trips to London .
4	staff-pos clean-pos bed-pos	The room was quiet, clean, the bed and pillows were comfortable, and the service was
5	readiness-pos	We arrived about 11 am, room was ready .
6	size-pos clean-pos	Room was good size for Europe , clean throughout .
7	staff-pos	The Concierge desk called our room to ask if we needed any information or assistance .
8	size-pos clean-pos bed-pos	Room was plenty big enough and clean and tidy, bed was comfordable .
9	equip-neg	First, we walked in and the restroom door was broken .
10	clean-pos	Our room was typical holiday inn the bathroom could have done with updating but was
11	readiness-neg	Our rooms were not ready, we were promised rooms at a later time, etc .
12	size-pos	My room was positively huge by European standards .

Processing description

- dividing the data set into the learning (350 opinions) and the testing sets (33 opinions)
- for the learning set (containing room descriptions and labels):
  - stemming
  - usage of stop-list filter
  - Labeled LDA model building
- for testing set (containing only room description):
  - label prediction
  - model evaluation

Algorithm implementation



```
val source = CSVFile("hotel-rooms-learn.c

val tokenizer = {
  SimpleEnglishTokenizer() ~>
  CaseFolder() ~>
  WordsAndNumbersOnlyFilter() ~>
  MinimumLengthFilter(3)
}

val text = {
  source ~>
  Column(3) ~>
  TokenizeWith(tokenizer) ~>
  TermCounter() ~>
  TermMinimumDocumentCountFilter(3) ~>
  //TermDynamicStopListFilter(30)// ~>
  DocumentMinimumLengthFilter(3) ~>
  //StopWordFilter("en")
  TermStopListFilter(List(
  "most","and","can","was","had","with","
  "kept","going","out","wasn't","what","p
  "other","did","even","throughout","etc"
  "175","maybe","150","around","that's","
  "itself","then","being","said","your","
  ))
```

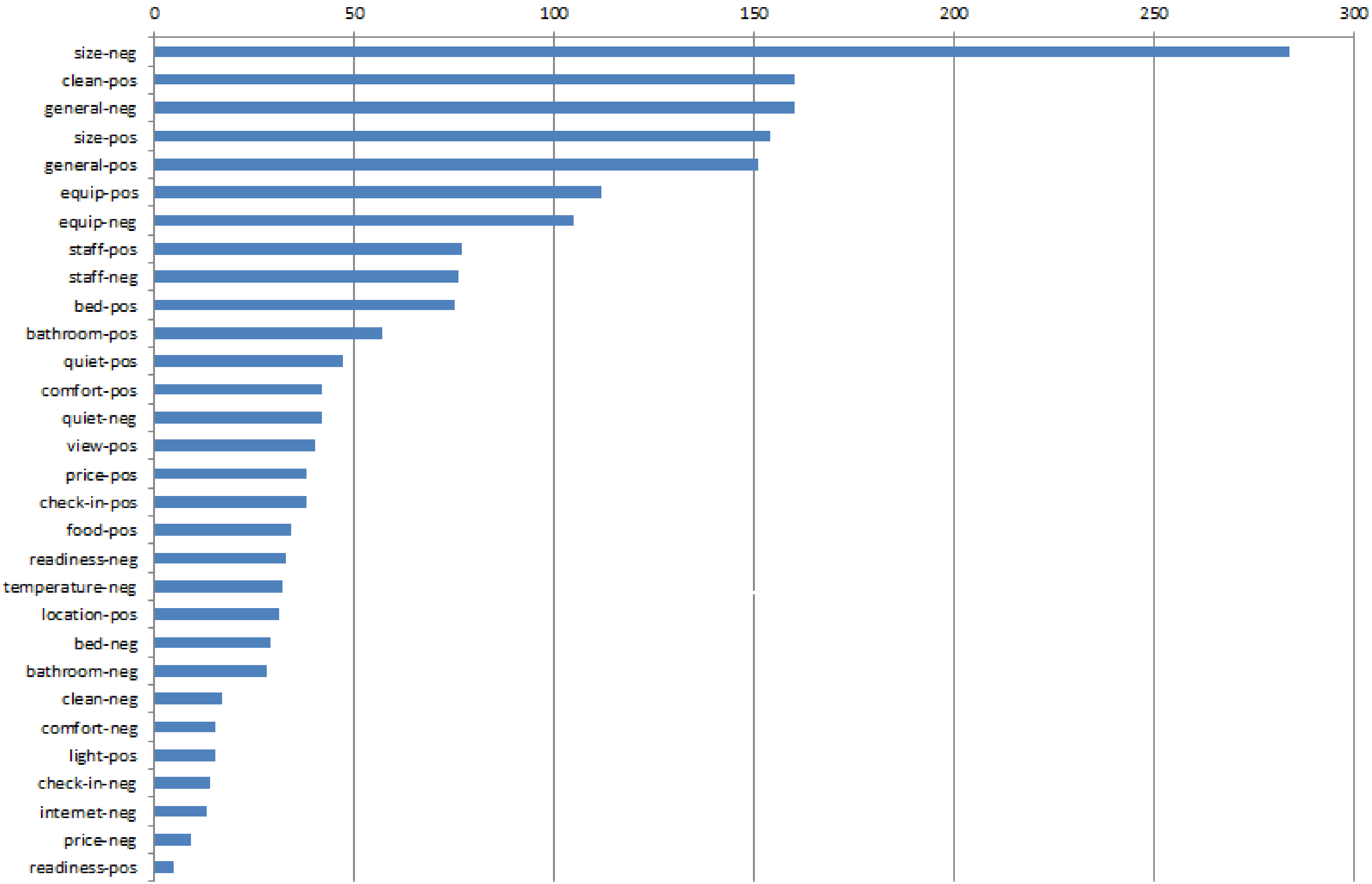
Stanford Topic Modeling Toolbox

Load a TMT script into a new tab using the File -> Open script.

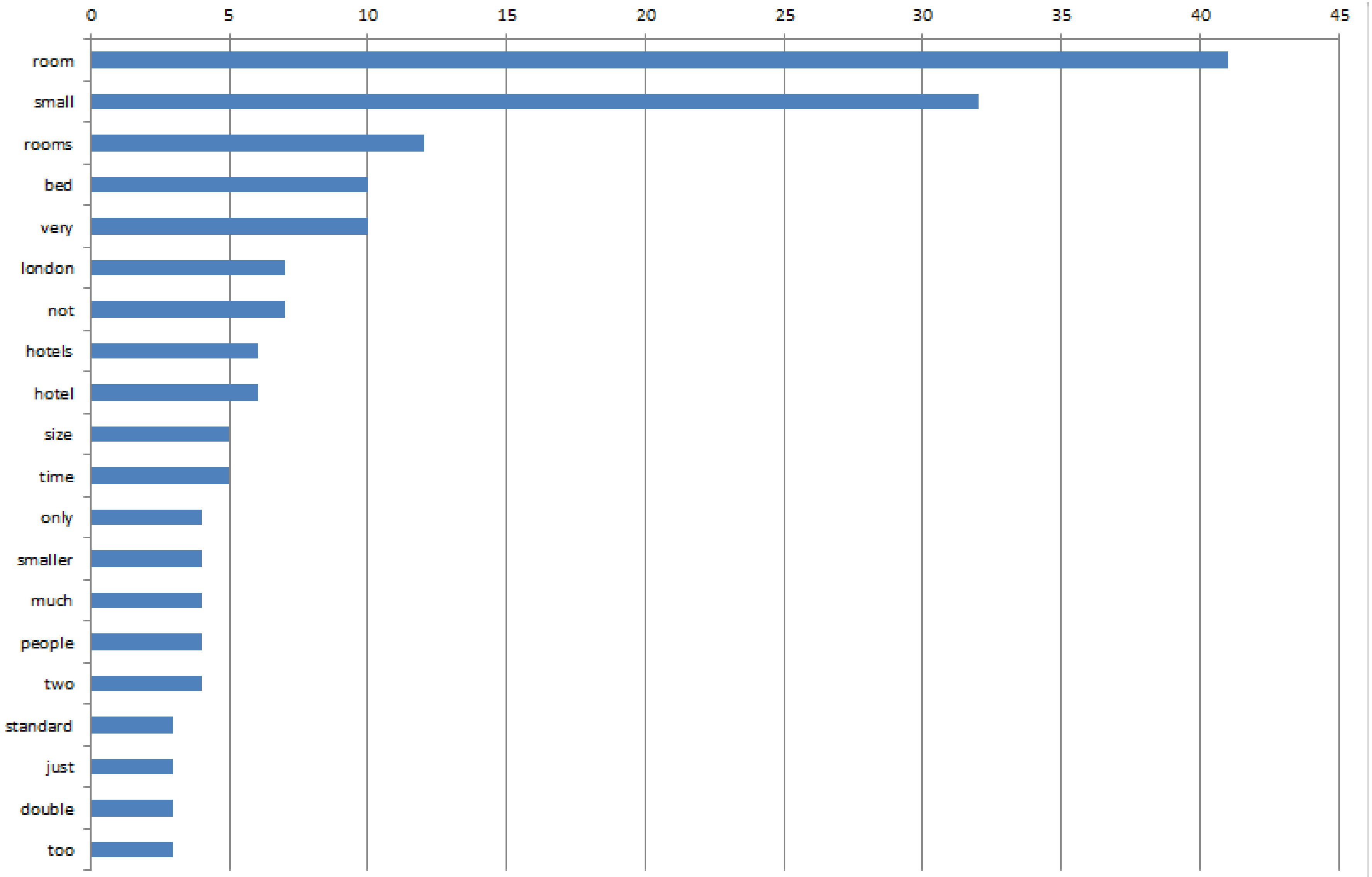
Copyright (c) 2009- The Board of Trustees of
The Leland Stanford Junior University. All Rights Reserved.

NLP package for Scala language and Stanford Topic Modeling Toolbox.

Topics and their importance



Topic: size-neg



Prediction example (for testing set)

Opinion:

The bathroom is a good size .

Labels:

bathroom-pos

Model results:

bathroom-pos (1,0)

Prediction example (for testing set)

Opinion:

The room was clean and, by London standards, decently sized .

Labels:

clean-pos

size-pos

Model results:

size-pos (0,98),

clean-pos (0,02)

Prediction example (for testing set)

Opinion:

When we tried to use a phone card from our room it would not work so I asked the front desk to help me and was told they couldn't really !

Labels:

staff-neg

Model results:

staff-neg (1,00)

Prediction example (for testing set)

Opinion:

The hotel room was very clean and the cleaning staff and breakfast staff were very attentive .

Labels:

clean-pos

staff-pos

Model results:

staff-pos (0,7)

clean-pos (0,3)

Model evaluation

- IR measures (for testing set):
  - Precision = 0,94
  - Recall = 0,98
- Advantages:
  - high quality
- Disadvantages:
  - time-consuming process of model building

Thank you for your attention!

OPINION MINING AND ITS IMPORTANCE IN THE ACTIVITY OF CONTEMPORARY ENTERPRISES