INFOX
educación sin fronteras

Text Mining – Sentiment Analysis
Visualización de datos

Mg. Gloria Rivas

Agenda

1. **Sentiment and opinions**
2. **Cuantificando las emociones**
3. **Analizando las oraciones**
4. **What to do with sentiment?**

INFOX
educación sin fronteras

Agenda

1. Sentiment and opinions

Sentiment and Opinions (Merriam – Webster)

Sentiment

Una actitud, pensamiento o juicio provocados por sentimientos.

- Sentimientos reflejando una emoción
- Ejemplo: “I worry about the healthiness of the air in this hotel”

Opinion

Una opinión, juicio o valoración formada en la mente sobre un asunto en particular.

- Un concreto punto de una persona hacia algo
- Ejemplo: “I think the air in this hotel is unhealthy”

Sentiment analysis

- Los comentarios están más cercas a las opiniones que a los sentimientos
- Pero el análisis es usualmente llamado *sentiment analysis* (o *opinion mining*)

Por qué es importante entender los sentimientos?

Las opiniones son claves para influenciar el comportamiento.

- Creencias y percepciones de la realidad son construidas sobre cómo otros ven el mundo.
- Siempre que tomamos una decisión, usualmente buscamos el input de los otros
- El incremento de las redes sociales incrementa la disponibilidad de la data de opinión

❑ Para consumidores
❑ Para investigación relacionada al marketing

- Preguntas importantes

❑ Mide el sentimiento
❑ Evalúa el sentimiento sobre el tiempo
❑ Identifica las causas del sentimiento

También:
Importante para desarrollar interacciones automatizadas como chatbots (pero en este curso no veremos eso)

Sentiment mining (basado en Wikipedia)

- Opinion mining (aka sentiment mining) usa programas de computadora para sistemáticamente

✓ Identificar
✓ Extraer
✓ Cuantificar, y
✓ Estudiar

Text Analytics

Estados afectivos e información subjetiva

- Sentiment mining es usualmente aplicado a

✓ Voces del consumidor: reviews, survey responses
✓ Online and social media

- Healthcare materials

Sentiment mining (basado en Wikipedia)

- Sentiment analysis busca
- Determinar la actitud
- Con respecto a cierto tema

O una reacción a
- Un documento
- Interacción
- O evento

La actitud puede ser
- Un juicio o evaluación
- Estado afectivo, o
- Una comunicación emocional intencionada

Sentiment

Una opinión

1. De una persona
2. Sobre un objetivo (o aspecto de un objetivo)
3. Sobre una emoción
4. O una fuerza

Leading example: Sentiment analysis from reviews

An example review:

Even though the Travelodge is inexpensive, I would never return here. Serious noise problem due to the fact that it is in close proximity to city fire station. The sirens went off multiple times in the middle of the night. Also lots of outdoor noise from outdoors late at night. Also, hotel is remote - you have to walk for blocks to get to a restaurant. Lack of cleanliness was apparent – carpet was filthy - big stains. Front desk staff were not particularly helpful, but breakfast was nice.

What sentiment information (opinion) is present in this review?
→ Please try to summarize this for yourself

Sentiment analysis from reviews

An example review:
Even though the Travelodge is inexpensive, I would never return here. Serious noise problem due to the fact that it is in close proximity to city fire station. The sirens went off multiple times in the middle of the night. Also lots of outdoor noise from outdoors late at night. Also, hotel is remote - you have to walk for blocks to get to a restaurant. Lack of cleanliness was apparent – carpet was filthy - big stains. Front desk staff were not particularly helpful, but breakfast was nice.

Relevant questions:

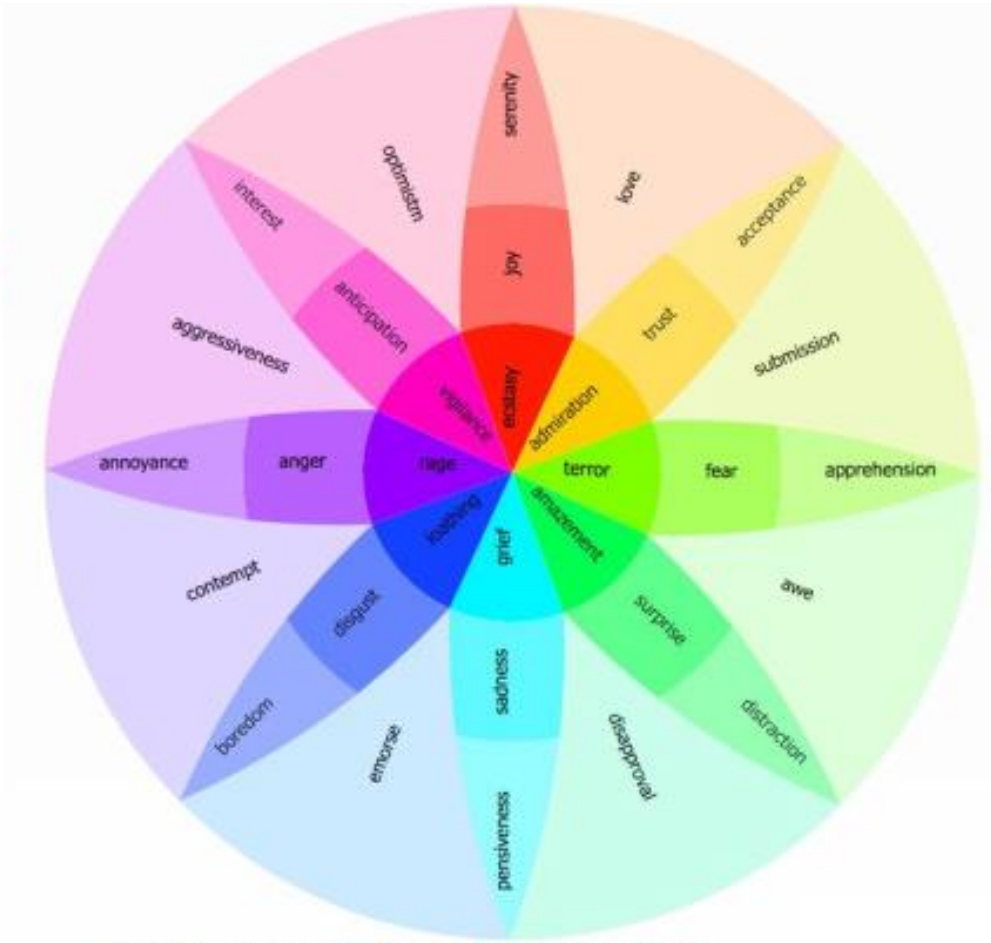
- Overall sentiment?
- What sentiment information is present?
- Which sentiments are positive / negative / neutral?
- Which sentiments are weak and which are strong?

Agenda

2. Cuantificando las emociones

Cuantificando los sentimientos

- Positive versus negative
- By emotion (nrc sentiment dictionary)
  - Anger
  - Fear
  - Anticipation
  - Trust
  - Surprise
  - Sadness
  - Joy
  - Disgust



Plutchik's wheel of emotion

Cuantificando (= contando) los sentimientos

Much of sentiment analysis is based on *valence of words*

- Negations are sometimes ignored
- Sarcasm is very challenging
  → Requires Natural Language Processing [NLP] algorithms (deep insight in sentences)

Most simple idea:

- Use a dictionary (a list) for positive (and negative) words
- For each word in the text check whether it appears in the list
- Count positive/negative words
- `syuzhet` package provides multiple approaches

→ Do not apply stemming! Why?

Sentiment dictionaries

Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, Aug 22-25, 2004, Seattle, Washington, USA.

1. Diccionario BING (6 789)
    - 4 783 negativas
    - 2006 positivas
2. Lista generada por investigadores
3. Empezó en el 2004
4. Sigue siendo actualizada

Ejemplo

`(get_sentiment_dictionary('bing'))`:

Negative words	Positive words
2-faced	a+
2-faces	abound
abnormal	abounds
abolish	abundance
abominable	abundant
abominably	accessable
	accessible

INFOX
educación sin fronteras

Sentiment dictionaries

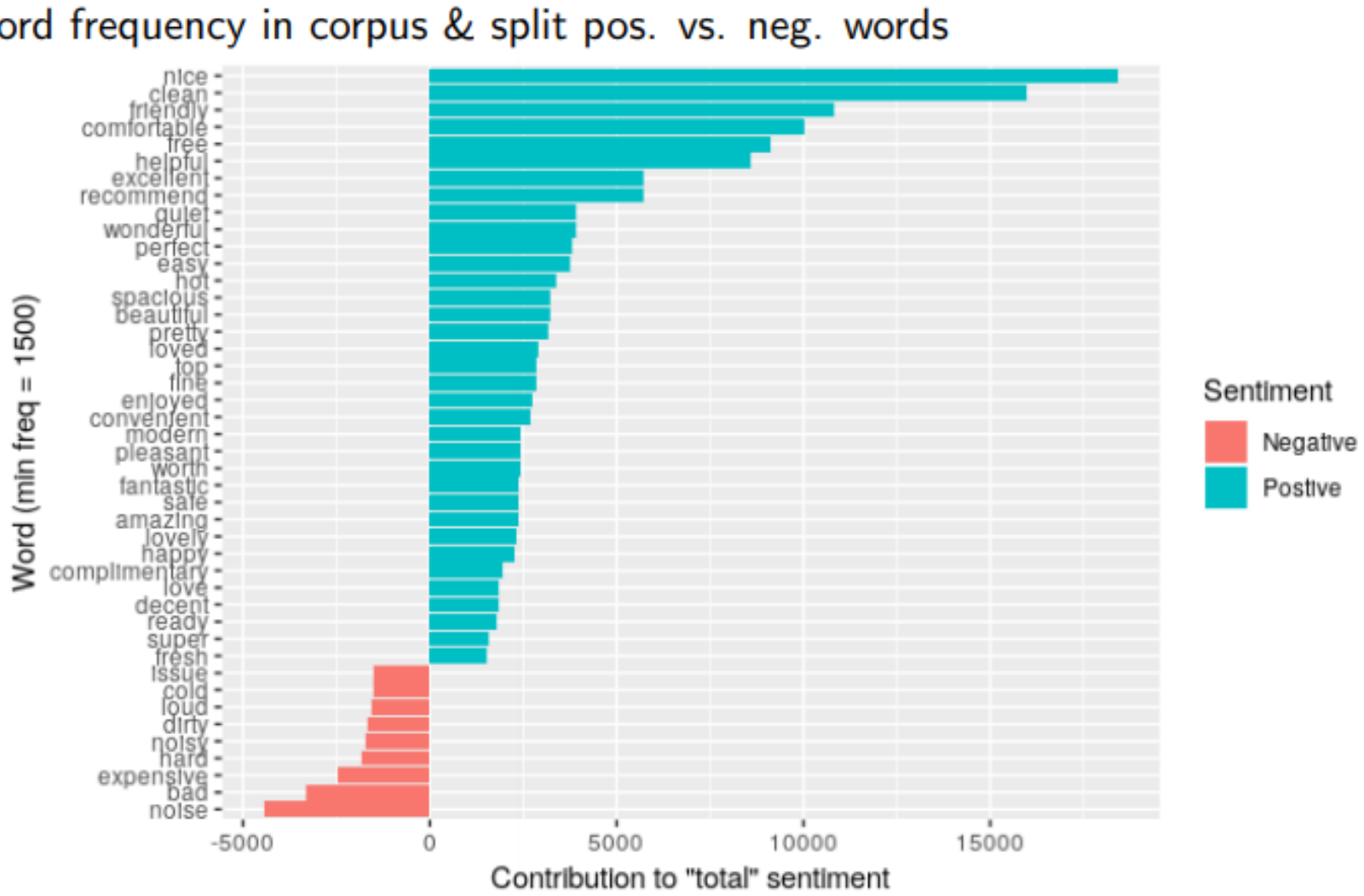
Saif Mohammad and Peter Turney. "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon." *In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, June 2010, LA, California.

1. NRC sentiment y un diccionario de emociones[National Research Council Canada]
2. Crowdsourced data collection
3. Fewer positive and negative words
4. But emotional valence coded as well

Ejemplo

word	emotion	flag
happy	anger	0
happy	anticipation	1
happy	disgust	0
happy	fear	0
happy	joy	1
happy	negative	0
happy	positive	1
happy	sadness	0
happy	surprise	0
happy	trust	1

Overall sentiment



INFOX
educación sin fronteras

Contando ‘sentiments’ en un review

- ❑ "the room was kind of clean but had a very strong smell of dogs." (1pt)
- ❑ "generally below average but ok for a overnight stay if you're not too fussy." (-1pt)
- ❑ "would consider staying again if the price was right." (1pt)
- ❑ "breakfast was free and just about better than nothing." (2pt)

Total: 3 puntos

Notas:

- ❖ Las palabras no están tomando en cuenta el contexto.
- ❖ En general estas oraciones no son muy positivas y un poco sarcásticas.

Qué pasa con las negaciones y las amplificaciones?

Ejemplo: "The food is very delicious, but the service is not so good“
Tendría una puntuación 2

Ajustamos por negación, amplificaciones y largo de la oración.

- Encontramos la palabra que aparece en el diccionario (= polarity word)
- Consideramos 4 palabras anteriores y las 2 siguientes palabras
- Polarity word positive =+1 (negativo -1)
- Cambiamos el signo si palabra es negativa
- Añadimos y substraemos 0.8 por cada amplificador
- Sumamos las polarity words
- Dividimos entre $\sqrt{\text{no. words}}$

Usando la función de polaridad, la oración toma una puntuación de 0.23.

INFOX
educación sin fronteras

Qué pasa con las negaciones y las amplificaciones?

"The food is very delicious, but the service is not so good"

"The food is very delicious, but the service is not so good"

1

1

"The food is very delicious, but the service is not so good"

0.8 1 -

1

$$\frac{0.8}{\sqrt{12}} = 0.23$$

INFOX
educación sin fronteras

Qué pasa con las negaciones y las amplificaciones?

Para practicar:

- ❑ ”The food is good”
- ❑ ”The food is very good”
- ❑ ”The food is bad”
- ❑ “The food is not bad”
- ❑ “The food is not very good”

Respuestas: 1, 1.8, -1,1,-0.2

INFOX
educación sin fronteras

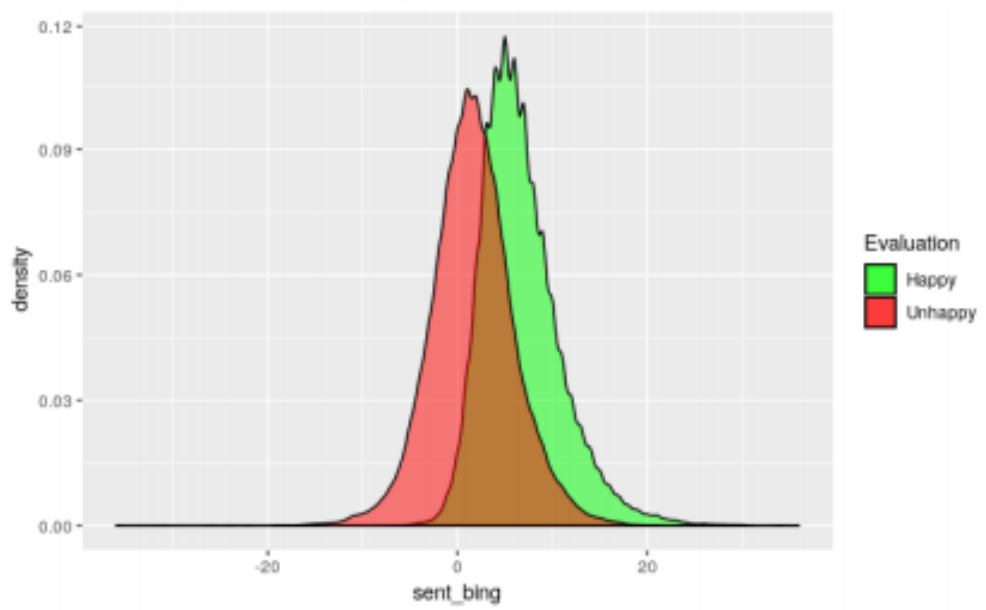
Agenda

3. Analizando las oraciones

Does sentiment matter?

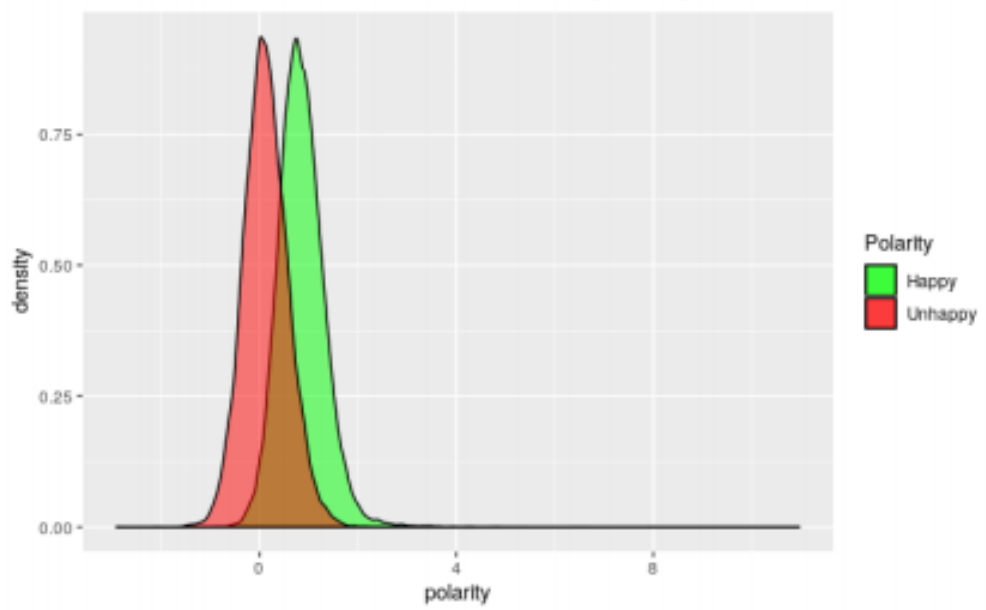
Does sentiment score relate to overall rating (happy/unhappy)?

Bing dictionary



Mean = 1.61 (unhappy) vs mean = 6.70 (happy)

Polarity algorithm



0.13 (unhappy) vs 0.85 (happy)

Zooming in

- We already knew what the positive and negative reviews focused on
- Sentiments by review do not add much relative to the `happy/unhappy` score

How can we get more insight in what reviewers complain about or are happy about?

- "stayed here with husband and sons on the way to an alaska cruise." (0p)
- "we all loved the hotel, great experience." (2p)
- "ask for a room on the north tower, facing north west for the best views." (1p)
- "we had a high floor, with a stunning view of the needle, the city, and even the cruise ships!" (1p)
- "we ordered room service for dinner so we could enjoy the perfect views." (2p)

→ Look at words in "important" sentences

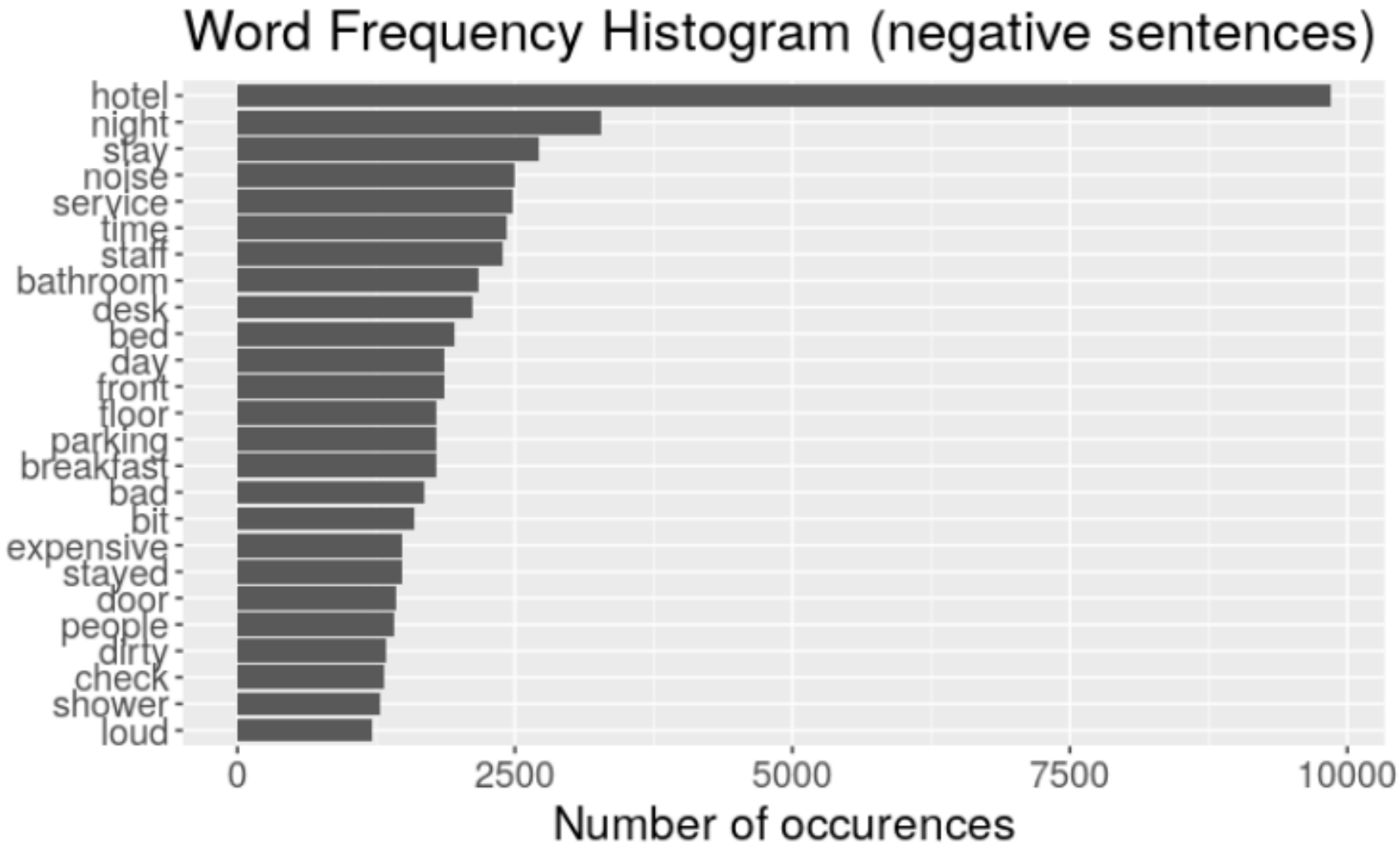
Zooming in

Asking: *What words are mentioned in positive reviews*

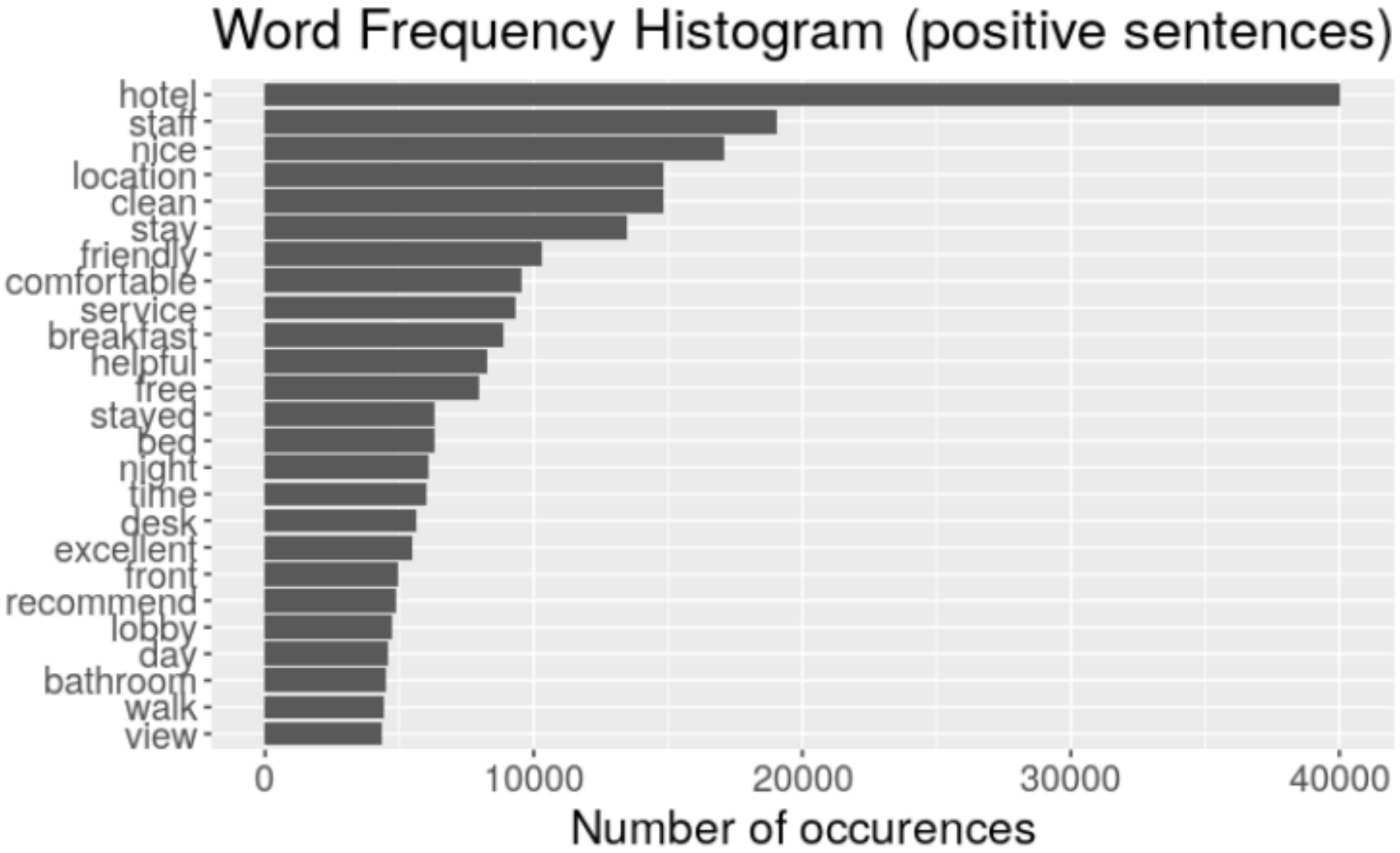
versus

Asking: *What words are mentioned in positive sentences* &
What words are mentioned in negative sentences

Zooming in negative sentences



Zooming in positive sentences



Las personas de qué se quejan?

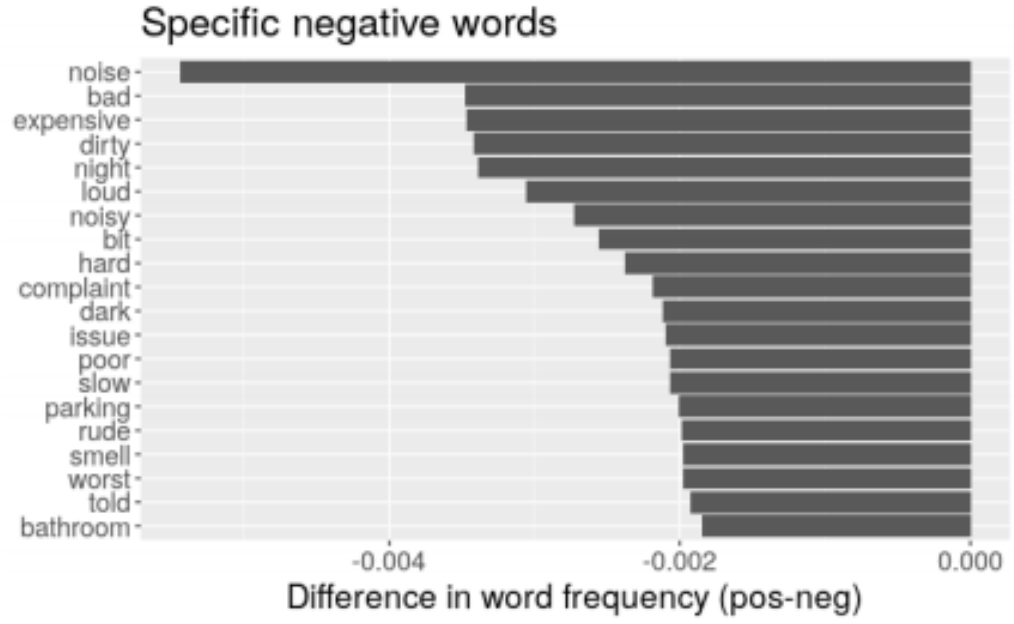
With this approach:
→ some standard words show up
→ *negative words* appear more often by definition

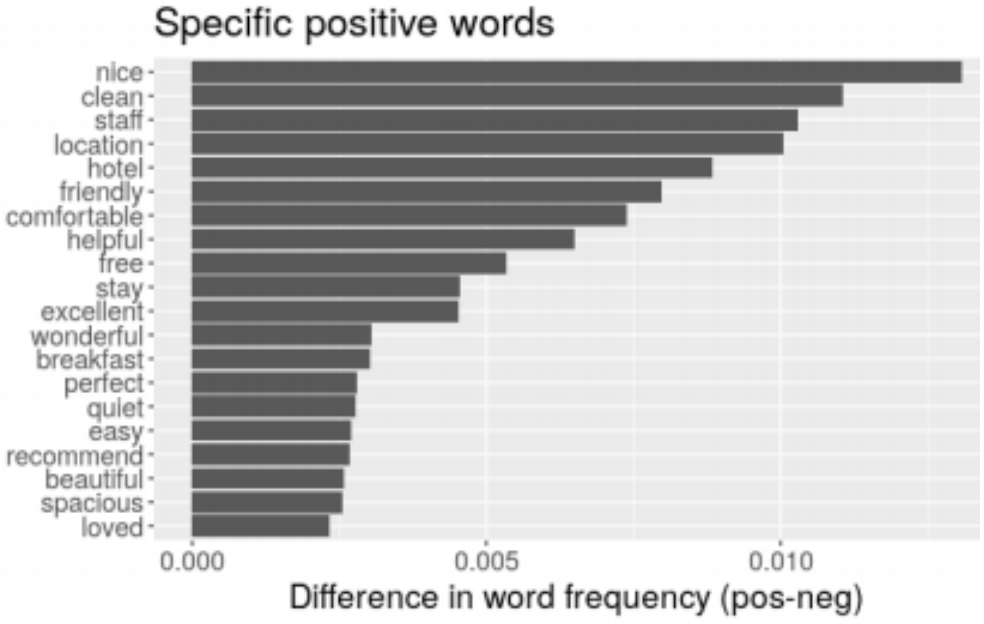
Question 1: Is the word "bad" informative (by itself)?

Question 2: Is the word "hotel" informative (by itself)?

Look at the differences!

Frequency is not very informative! → differences in (or ratios of) word frequencies are





INFOX
educación sin fronteras

Type of word

Look at type of word:

→ What type of words could be most useful to determine what they complain about?

- Verbs?
- Nouns?
- Adjectives?

Selecting nouns of sentence

Recognize type of words in a sentence:

- *Part-of-Speech [POS] tagging*
- Is complex task!

Multiple algorithms exist

- RDRPOSTagger uses *Ripple down rules* (a decision tree like method)

Type of words

Some types of words (see this link for full list)

Nr.	Code	Meaning
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
7.	JJ	Adjective
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present

Part of speech tagging

Analyze: *"The room was kind of clean but had a very strong smell of dogs."*

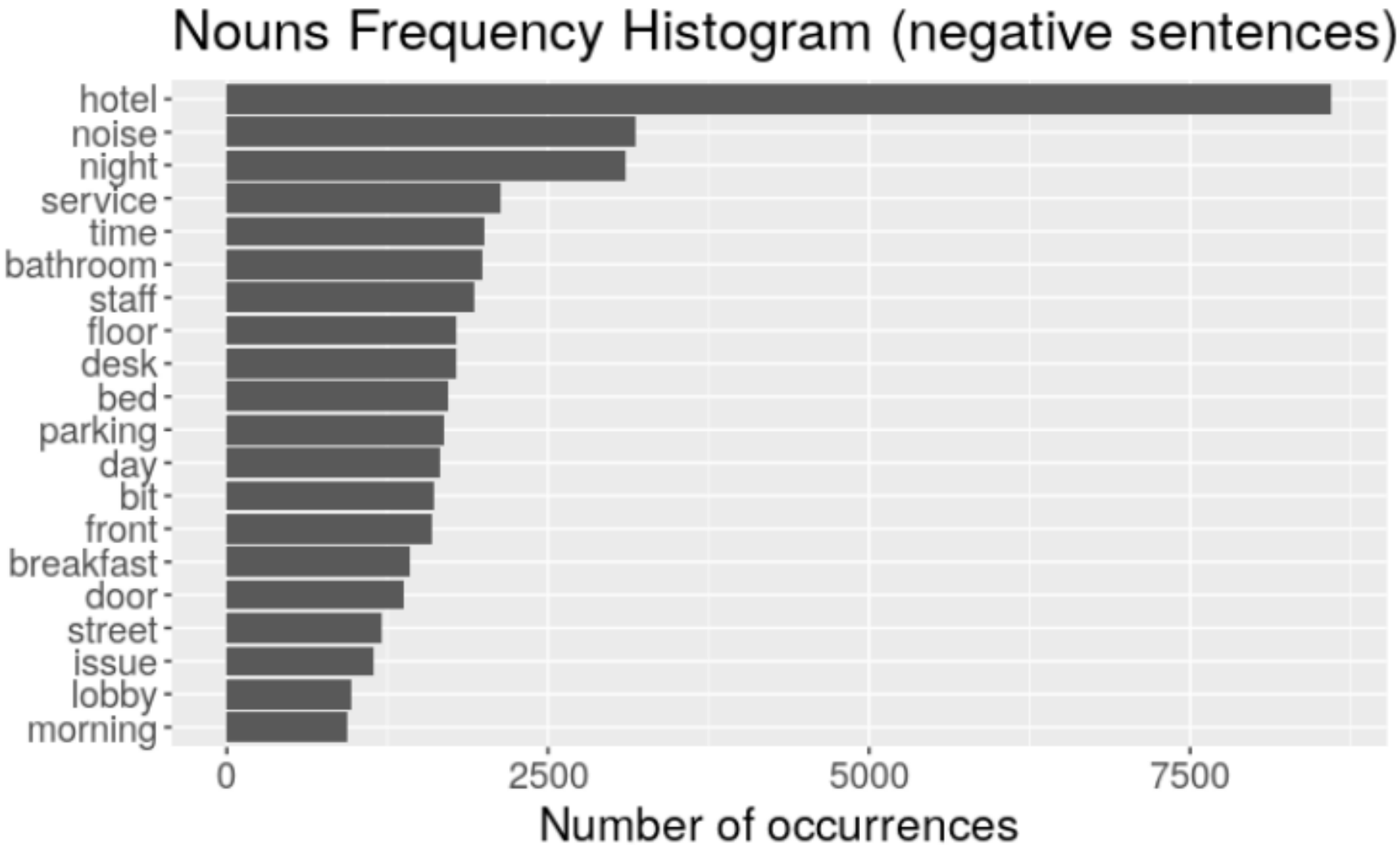
```
>rdr_pos(POS_specs,"The room was kind of clean but had a very strong smell of dogs.")
doc_id token_id  token pos
```

doc_id	token_id	token	pos
1	d1	1	The	DT
2	d1	2	room	NN
3	d1	3	was	VBD
4	d1	4	kind	NN
5	d1	5	of	IN
6	d1	6	clean	JJ
7	d1	7	but	CC
8	d1	8	had	VBD
9	d1	9	a	DT
10	d1	10	very	RB
11	d1	11	strong	JJ
12	d1	12	smell	NN
13	d1	13	of	IN
14	d1	14	dogs	NNS
15	d1	15	.	.

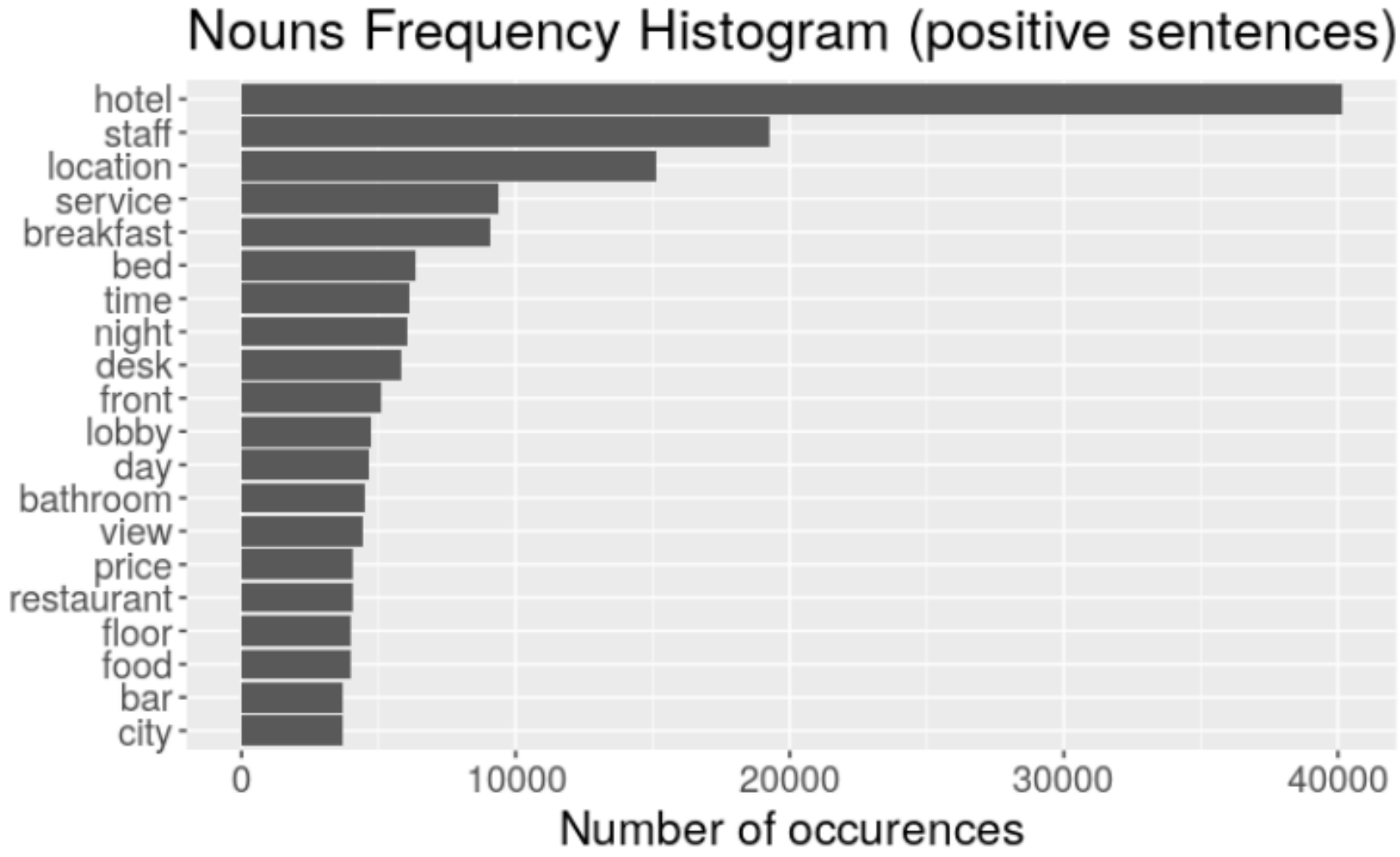
Notes on tagging

- Algorithms are not perfect!
- Function can also fail on malformed sentences (which are present in reviews!)
  $\rightarrow$ Pre-processing is important
- Capture errors using `tryCatch()`

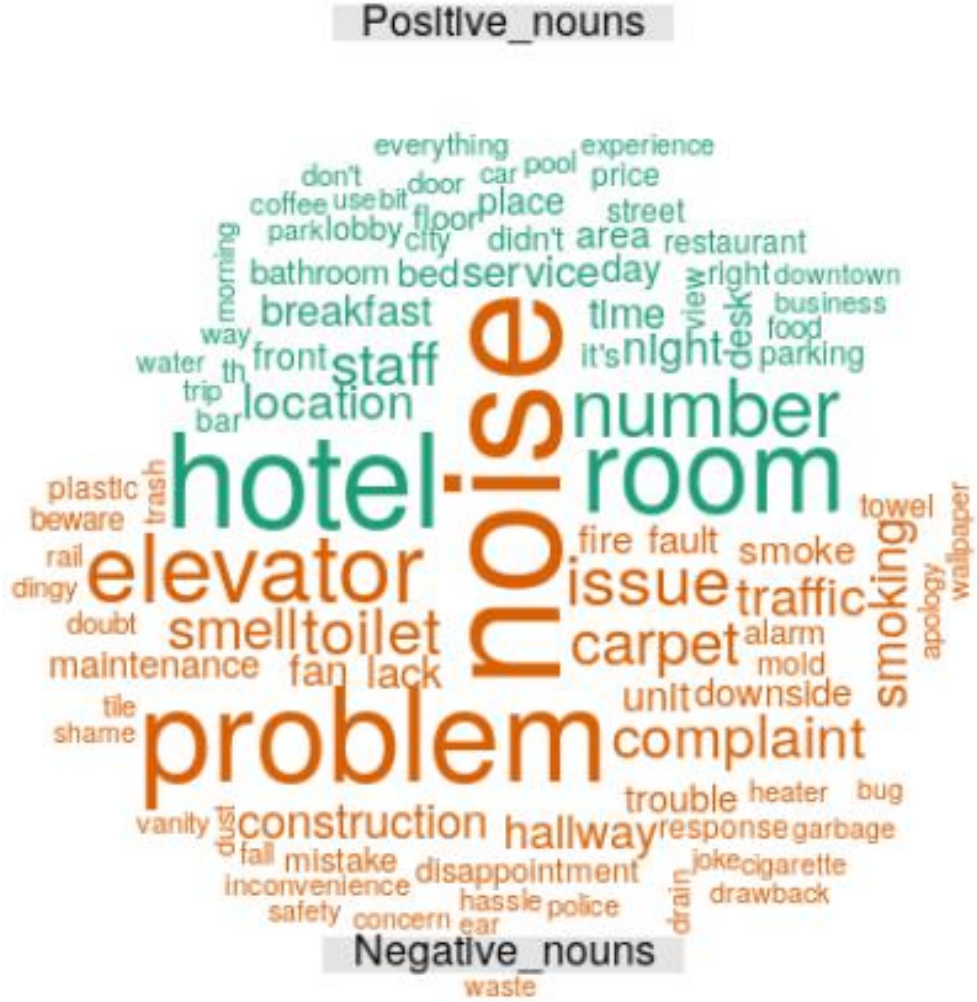
Nouns in negative sentences (only NN)



And for the positive sentences (only NN)



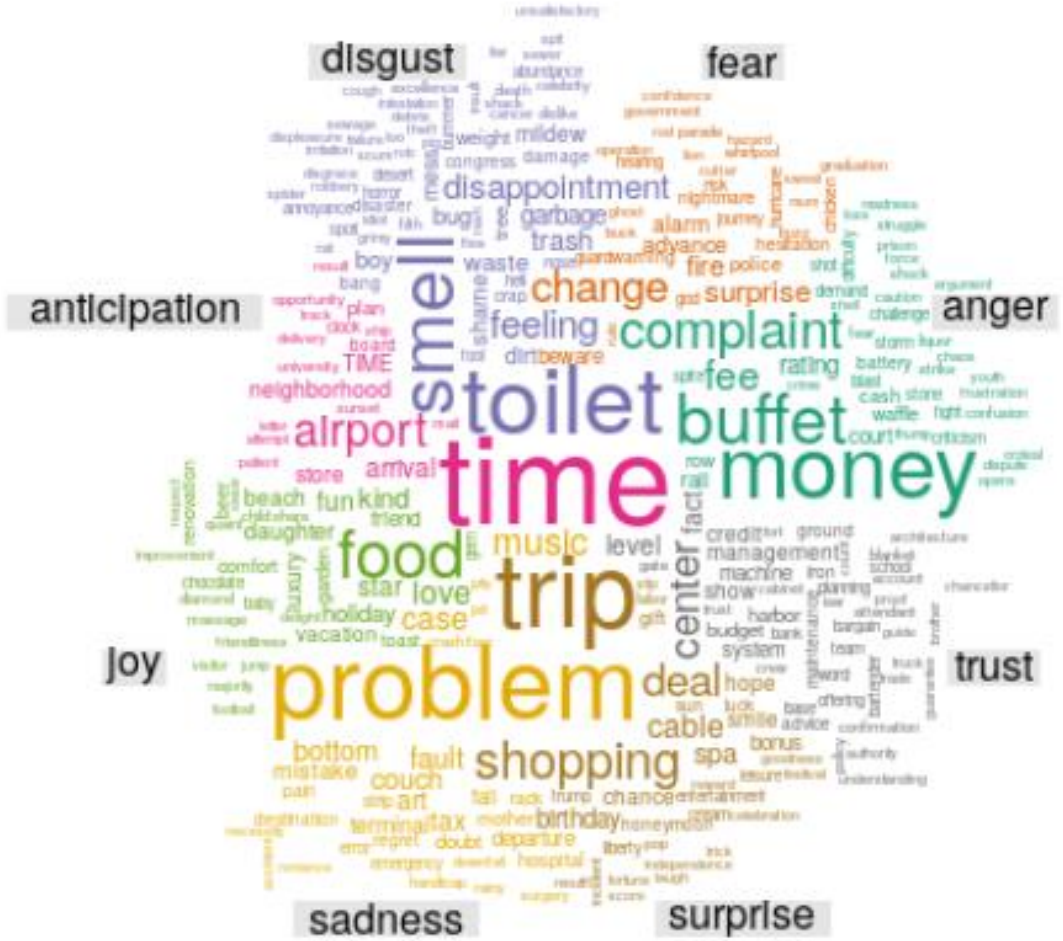
Visualizando usando comparison clouds



INFOX
educación sin fronteras

Podemos hacer lo mismo con las emociones

Find emotions using `get_nrc_sentiment()`



INFOX
educación sin fronteras

Problemas con las redes sociales

- Spelling errors: try to fix by replacing words, or spell checker
- Unexpected characters * # @ (keep, remove or replace)
- Terms like LOL: just keep (or replace)
- Emoticons: recognize and treat as word (or replace)
  → see book for details

Is this fun to listen to??

The actual fun, experience and learning is in doing this!

- Hands on experience in assignment and computer practicals
- Try other stuff yourself!
- Experiment $\rightarrow$ Search online $\rightarrow$ Solve

Agenda

4. What to do with sentiment?

Using sentiment analysis

Q: In which situations is sentiment analysis particularly helpful for firms?

When objective rating/scores are not available

- For many products still the case!
- Decision to enter an industry
- Investors choosing a stock to invest in
- Firms responding to a crisis

Samples vs population

Important question for sentiment analysis:
→ What is the source of your data?

Twitter, Facebook, Review websites,...
→ Not a representative sample of whole population!

Possible biases

- Age?
- Education?
- Wealth?
- ...
- Dissatisfied people?
- ...
- Trolling/Bots?

Samples vs population

Having the wrong sample frame → Wrong findings
→ Sample size cannot fix this!

What we can do

- Treat the data as giving *signals*
- If many people are dissatisfied about something → investigate it further
- Detect 'disasters'

Other topics

- Link sentiment to other variables (eg. demographics)
- Use inferred sentiment to predict something else. Rating/Purchase frequency/Return rate/...
- Tracking sentiment over time
  (and link to share/stock value/...)

¡GRACIAS!

INFOX
educación sin fronteras