INFOX educación sin fronteras

Clase 7 – Getting ready for Text Analytics

Mg. Gloria Rivas

Agenda

1. Text as data (Text is very different)
2. Cleaning the data (It starts over dirty)
3. Codification of text (Define language on words)

Los textos no son estructurados

Few familiar methods make sense

Analysis often more exploratory than explanatory

Can be linked to other, more structured data

- What features of a review lead to higher star-ratings?
- Which (type of) reviewers gives relatively more attention to a given set of attributes?
- Are there patterns within a reviewers behavior?

Cómo los analizamos?

ID	Browser	Device	Response	Description
id10326	Edge	Mobile	not happy	The room was kind of clean but had a VERY strong smell of dogs. Generally below average but ok for a overnight stay if you're not too fussy. Would consider staying again if the price was right. Breakfast was free and just about better than nothing.
id10329	Internet Explorer	Desktop	happy	Stayed here with husband and sons on the way to an Alaska Cruise. We all loved the hotel, great experience. Ask for a room on the North tower, facing north west for the best views. We had a high floor, with a stunning view of the needle, the city, and even the cruise ships! We ordered room service for dinner so we could enjoy the perfect views. Room service dinners were delicious, too! You are in a perfect spot to walk everywhere, so enjoy the city. Almost forgot- Heavenly beds were heavenly, too!
id10327	Internet Explorer	Mobile	not happy	I stayed at the Crown Plaza April -- - April --, ----. The staff was friendly and attentive. The elevators are tiny (about -' by -'). The food in the restaurant was delicious but priced a little on the high side. Of course this is Washington DC. There is no pool and little for children to do. My room on the fifth floor had two comfortable beds and plenty of space for one person. The TV is a little small by todays standards with a limited number of channels. There was a small bit of mold in the bathtub area that could have been removed with a little bleach. It appeared the carpets were not vacummed every day. I reported a light bulb was burned out. It was never replaced. Ice machines are on the odd numbered floors, but the one on my floor did not work. I encountered some staff in the elevator one evening and I mentioned the ice machine to them. Severel hours later a maid appeared at my door with ice and two mints. I'm not sure how they knew what room I was in. That was a little unnerving! I would stay here again for business, but would not come here on vacation.

Regression??

Correlation??

El texto como data

How can we code text data?

Let's try yourself (with your neighbor) and code text into data

"The room was kind of clean but had a VERY strong smell of dogs. Generally below average but ok for a overnight stay if you're not too fussy. Would consider staying again if the price was right. Breakfast was free and just about better than nothing."

At which level of detail did you code?

letter / word / combination of words / sentence / review

What did you code?

Presence of letter or word, use of CAPITALS, aspects mentioned, evaluation/attitude to aspect

El texto como data

How can we code text data?

What problems do you expect?

Can a computer do this?

First lecture is only on getting usable "data"

INFOX
educación sin fronteras

The word vector data model

Unit of observation

- Book
- Review
- Website
- Customer call

Provides one observation in data framework

- A row in excel or SPSS or R-dataframe
- Indicates whether a word is present (1) or not (0)

The word vector data model

- Ignores location and order of words

- This is a book that you do not want to put aside but instead read as soon as you can

- This is a book that you do not want to read but instead put aside as soon as you can

- Much can be learned even within this framework
  - Although it is not perfect

- Methods that account for sentence structure will also be covered
  - Word embeddings

Dataset de ejemplo

Data used is obtained from kaggle

- Many other datasets and codes available
- https://www.kaggle.com/anu0012/hotel-review

ID	Browser	Device	Response	Description
id10326	Edge	Mobile	not happy	The room was kind of clean but had a VERY strong smell of dogs. Generally below average but ok for a overnight stay if you're not too fussy. Would consider staying again if the price was right. Breakfast was free and just about better than nothing.
id10329	Internet Explorer	Desktop	happy	Stayed here with husband and sons on the way to an Alaska Cruise. We all loved the hotel, great experience. Ask for a room on the North tower, facing north west for the best views. We had a high floor, with a stunning view of the needle, the city, and even the cruise ships! We ordered room service for dinner so we could enjoy the perfect views. Room service dinners were delicious, too! You are in a perfect spot to walk everywhere, so enjoy the city. Almost forgot- Heavenly beds were heavenly, too!

38932 hotel reviews

6073063 words

The Word-vector data model

	The	room	was	kind	of	clean	but	had	a	VERY
Review 1	1	1	1	1	1	1	1	1	1	1
Review 2	1	1	0	0	1	0	0	1	1	0
Review 3	1	1	1	0	1	0	1	1	1	0

This is also called a Document-Term Matrix (DTM)

It's transpose is called a Term-Document Matrix (TDM)

INFOX
educación sin fronteras

Analizando la data con R

Coding in R:

<- assigns result on the right to variable on the left

%>% send the result to the next statement

```
reviews_df <- (read.csv("hotel-reviews.csv"))

print("number of reviews")
nrow(reviews_df)              38932

review_words <- reviews_df %>% unnest_tokens(word,Description)

print("number of words")
nrow(reviews_words)        6083298

counts <- review_words %>% count(word, sort=TRUE)
print("number of unique words")
nrow(counts)                  62940
```

INFO
educación sin fro

De qué se tratan los reviews?

```
#grepl returns true for the rows where 'regexp' is present, false otherwise
# How often is Chicago mentioned?
Chicago <- (grepl('Chicago', reviews_df$Description,ignore.case=T))
sum(Chicago)    1493

# How often is New York mentioned?
NY <- (grepl('New York', reviews_df$Description,ignore.case=T))
sum(NY)         2527

# How often is New York really mentioned?
NY <- (grepl(c('New York|NY'), reviews_df$Description,ignore.case=T))
sum(NY)          17216
```

INFOX
educación sin fronteras

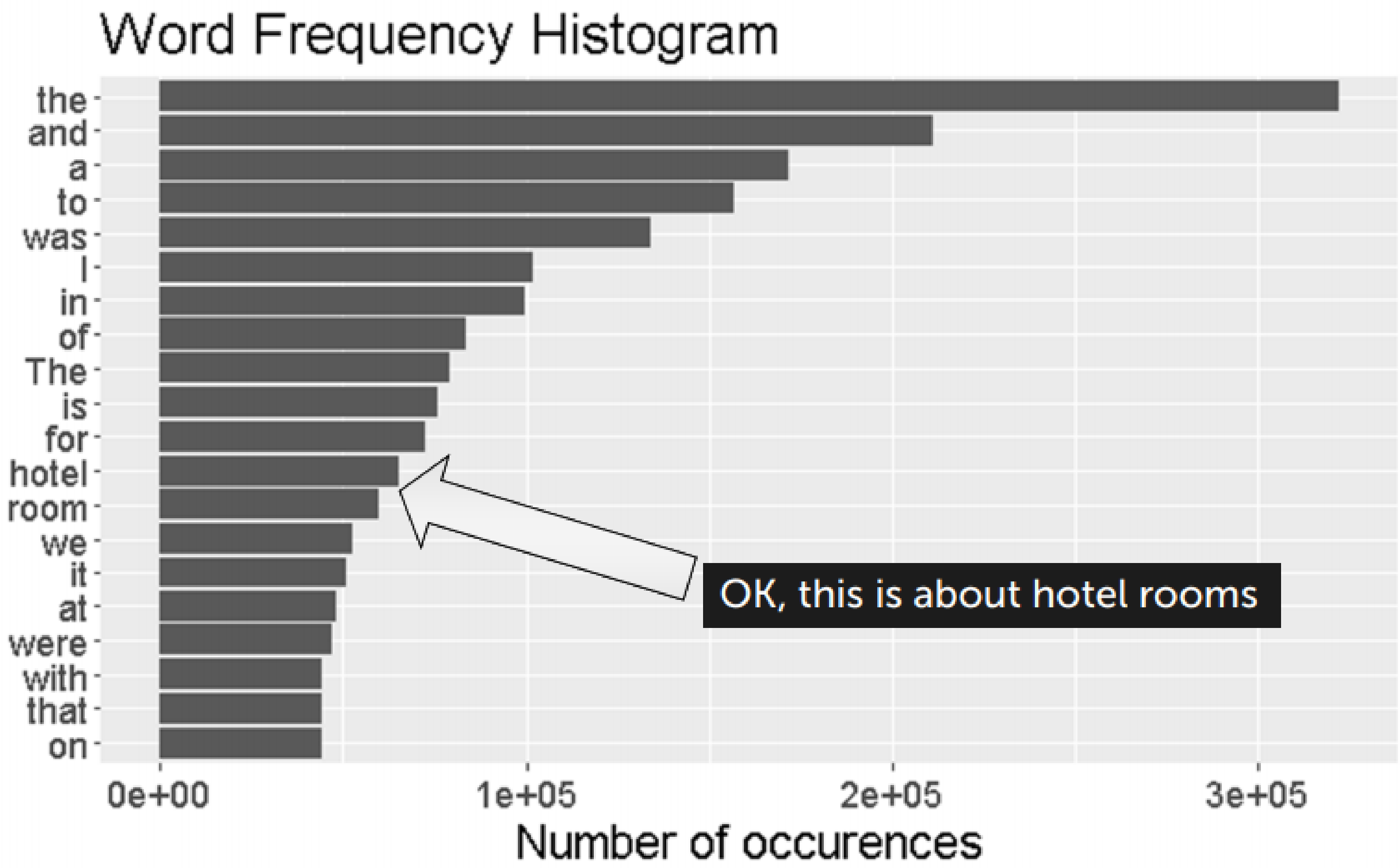
De qué se tratan los reviews?

```
#Let's see in which reviews the words 'Westin' AND 'New York' appear
Westin_in_NY <-
grepl(c('New York|NY'), reviews_df$Description,ignore.case=T)
& grepl('Westin', reviews_df$Description,ignore.case=T)
372

Westin_in_Ch <-
grepl(c('Chicago'), reviews_df$Description,ignore.case=T)
& grepl('Westin', reviews_df$Description,ignore.case=T)
48

#review share of Westin in Chicago and in NY
sum(Westin_in_Ch)/sum(Chicago)
sum(Westin_in_NY)/sum(NY)
.032
.022
```

INFOX
educación sin fronteras

The Word-vector data model

Data summary, most frequent words



INFOX
educación sin fronteras

La data es útil?

Stop words that are typically removed:

- A
- The
- One
- In
- No (!!!)
- Not (!!!)
- I
- We
- Have

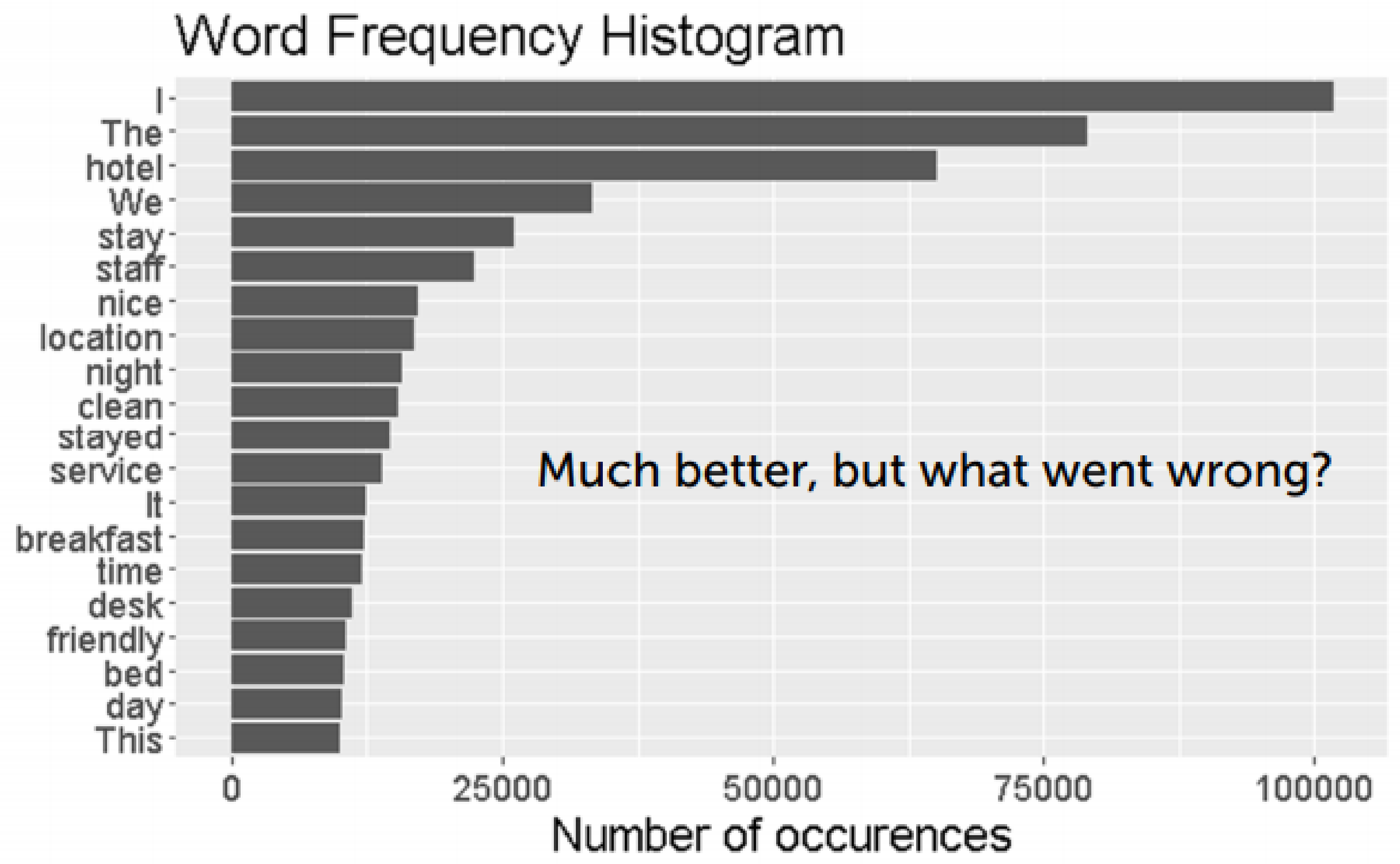
Data needs cleaning before starting an analysis

Garbage in is garbage out also holds for text analysis

Removiendo stopwords

How many from the 6083298 words are stop words?

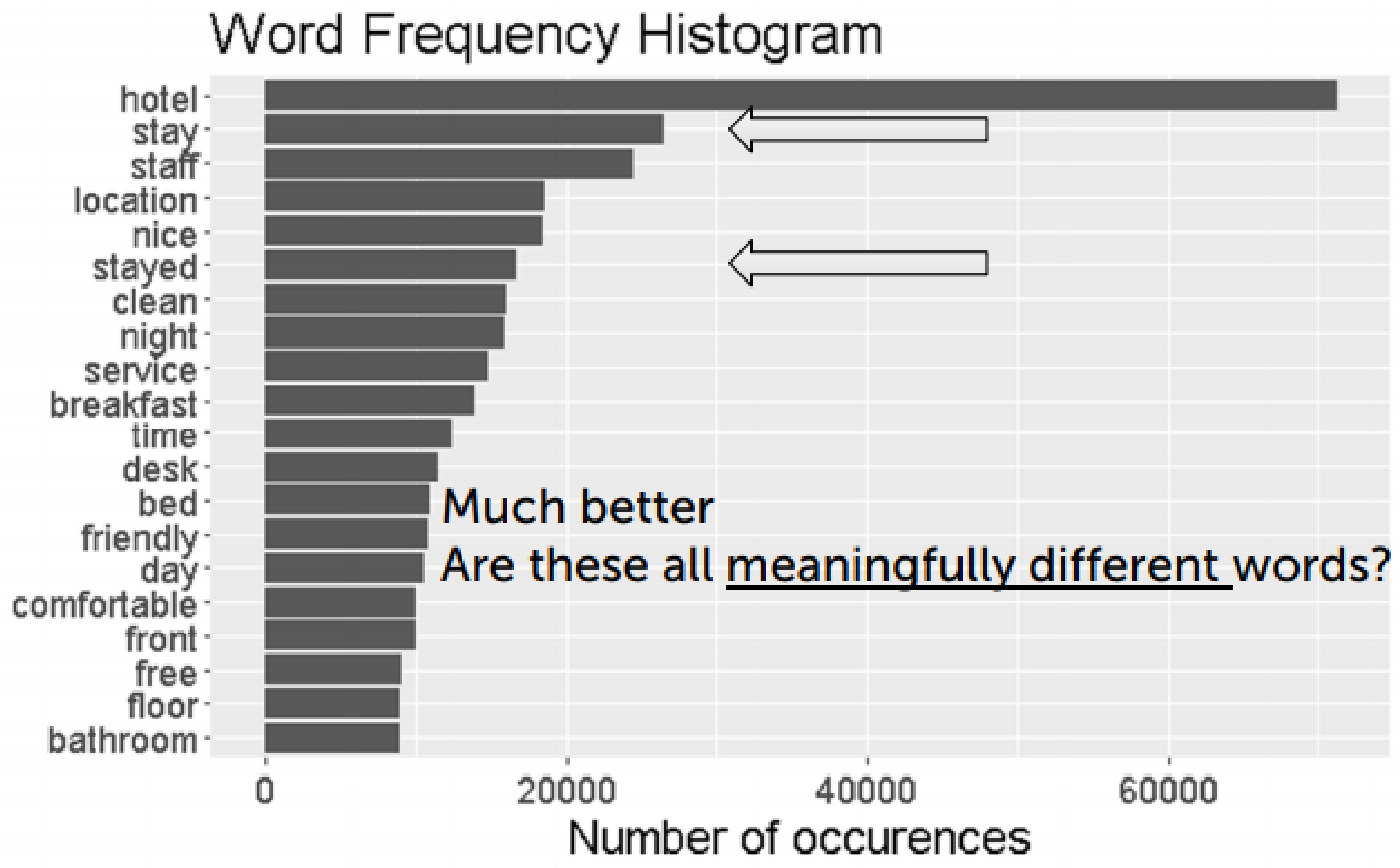
2165920 words remain, so 58% of words are stop words



Removiendo stopwords

How many from the 6083298 words are stop words?

2162163 words remain, so 65% of words are stop words



Diferentes palabras con significado similar

- For humans these are called synonyms
  - Love vs like
  - Poor vs bad
- Computers have a different definition of different
  - Love vs loves vs loving
  - Bad vs badly

Computers are stupid until we teach them something!

INFOX
educación sin fronteras

Diferentes palabras con significado similar

- Tell the computer what humans know to be similar words
- This is called stemming

Stem (definition):
A stem is the root or roots of a word, together with any derivational affixes, to which inflectional affixes are added.

Derivational affix -> change meaning of word
- Different meaning so different stem
  - Happy / unhappy / happiness

Inflectional affix -> change use of word but not meaning
- Same meaning so same stem
  - Happy / happily -> happy/i
  - Loves / loving -> lov

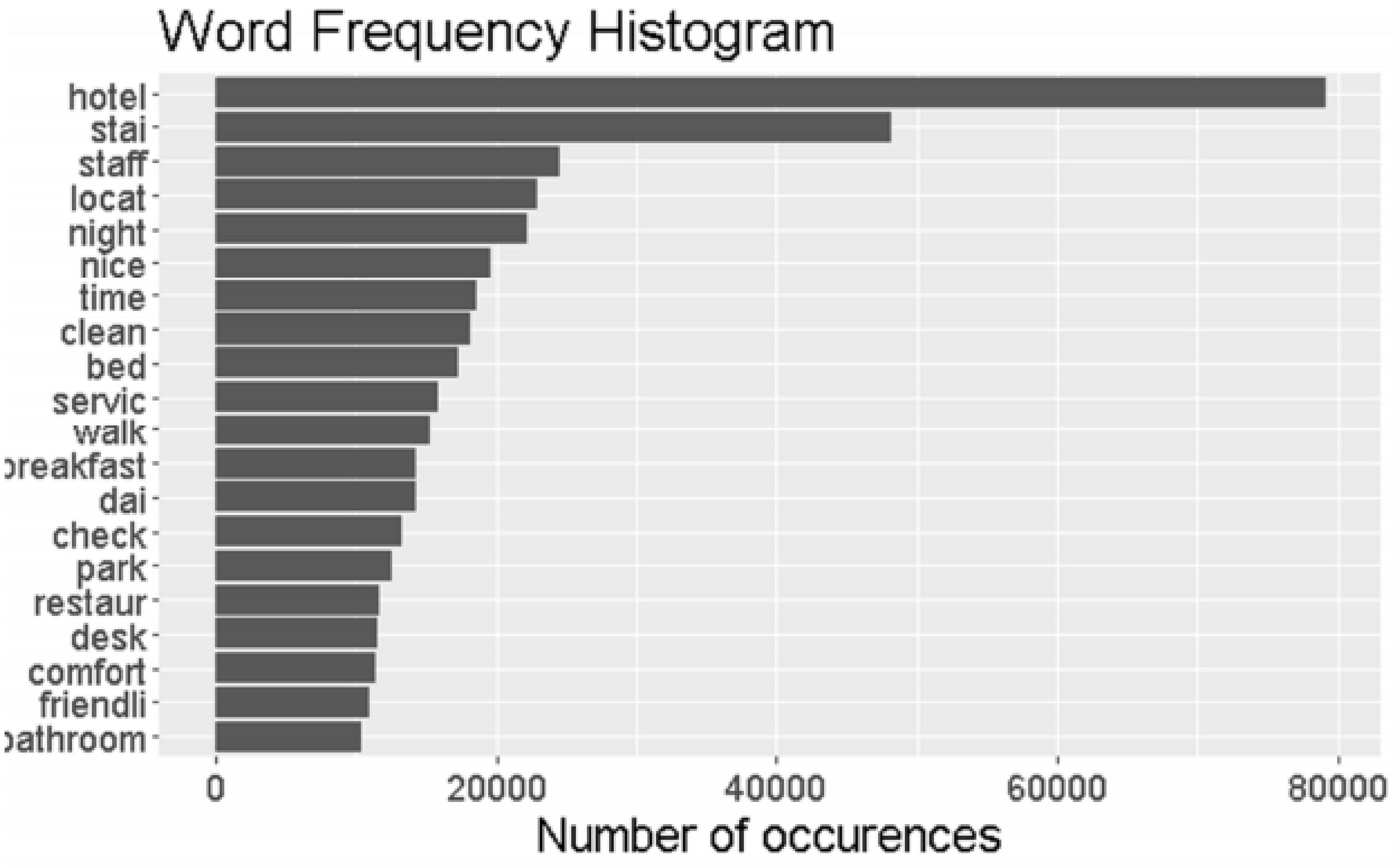
Stemming – Solo nos quedamos con la raiz

Famous stemming algorithm is by Porter

Removes common word endings
- ed
- ing
- able

But only when "allowed"
Has smart rules to differentiate between
    likable -> lik
    table -> table
    agreed -> agree
    bleed -> bleed
Based on coding of part that remains

Stemming – Solo nos quedamos con la raiz



Bed / beds
Clean / cleaned / cleaning
Stay / stayed / staying

INFOX
educación sin fronteras

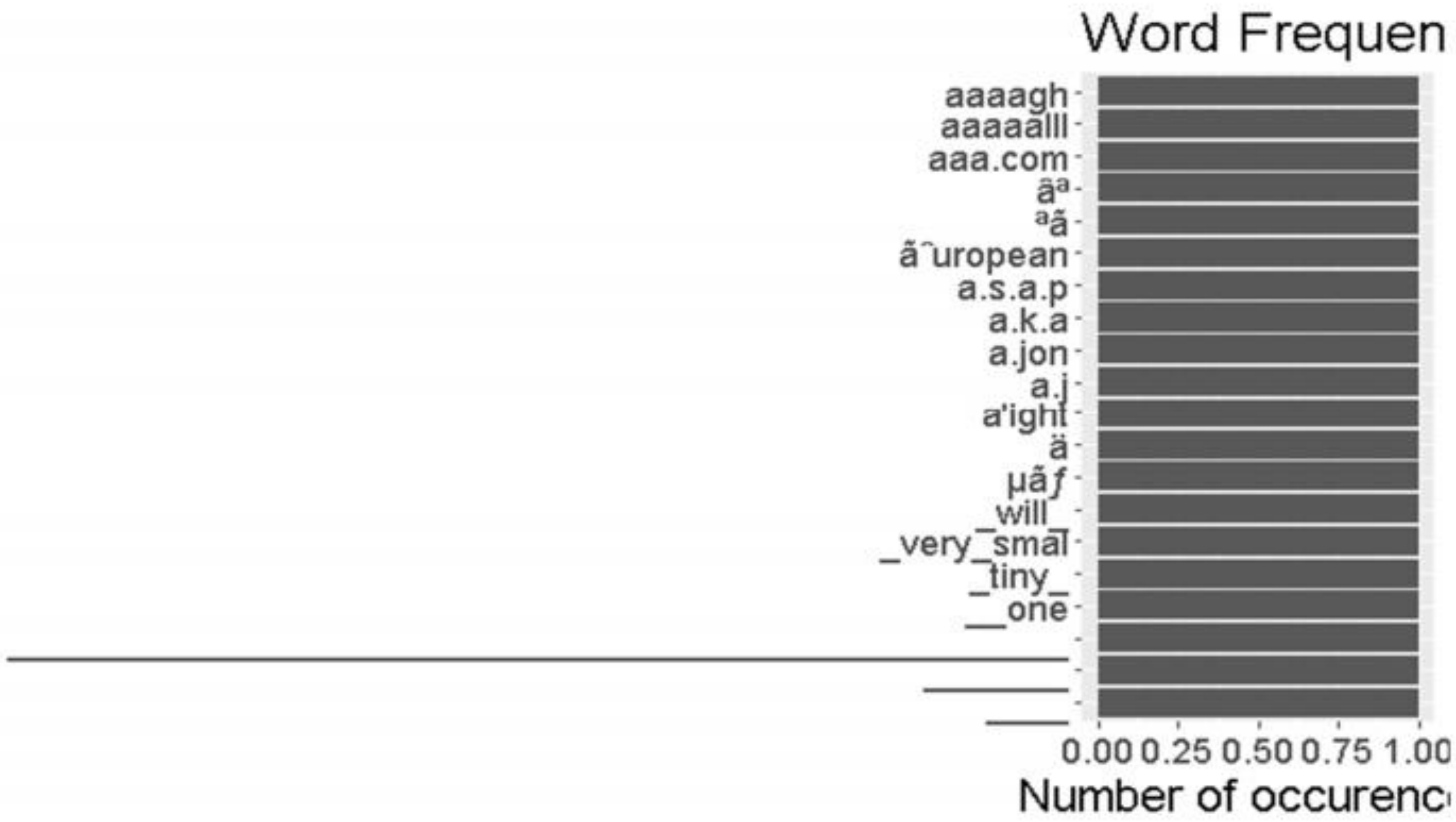
Insights de las palabras frecuentes/infrecuentes

What can we learn from words that occur extremely often?

What can we learn from a word that occurs only once?

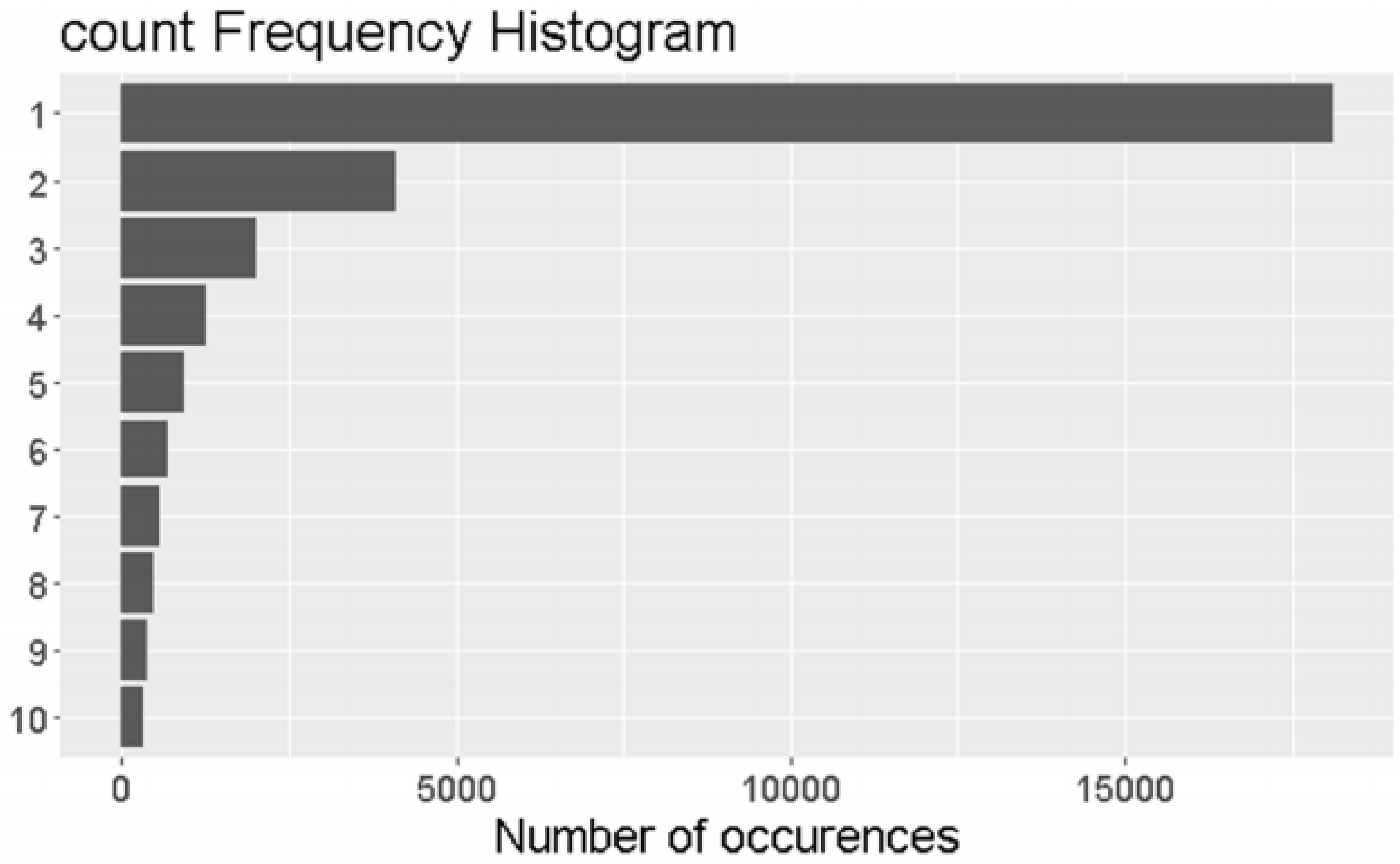
INFOX
educación sin fronteras

Palabras que sólo ocurren una vez



INFOX
educación sin fronteras

Distribución de la ocurrencia de la frecuencia de palabras



INFOX
educación sin fronteras

Focusing on useful variation

Remove the most frequent words

- hotel
- stai

Remove all 34990 words that occur fewer than (# reviews)/100 times

- cutoff is always a bit arbitrary

Focusing on useful variation

1+1=1?

Named entity extraction

Multiple words jointly have one meaning

Data science
Erasmus University
Air France
El Al
...

Use dictionary to integrate words into single unit
-> application specific

INFOX
educación sin fronteras

Resultado final de limpiar la base

- ✓ No stop words
- ✓ No capitals
- ✓ No punctuation
- ✓ Stemming applied

Strongly reduces size of dataset

fewer unique words

Screen out less informative words

very frequent words

very infrequent words

INFOX
educación sin fronteras

Volvemos a la base ...

Our hotel reviews are somewhat anonymized

- All numbers have been removed

Generally, numbers have little meaning beyond what they designate

Informative:

- Amounts of money
- Time of day

gsub(pattern , replacement, string) does global replacement of "pattern" with "replacement" in "string"

What does this command do?

gsub("\\$?-*\\.?-+" , " DOLLARVALUE " , reviewtext)

INFOX
educación sin fronteras

Visualización

INFOX
educación sin fronteras

Características de la visualización

- Palabras en un gráfico pueden tener

- ❑ Tamaño
- ❑ Locación
- ❑ Color
- ❑ Conexión

- Qué interferirías naturalmente de cada espacio?

- Después de haber limpiado la data, vamos a ver
- Word Frecuency
- Word Contrasts
- Word Similarity

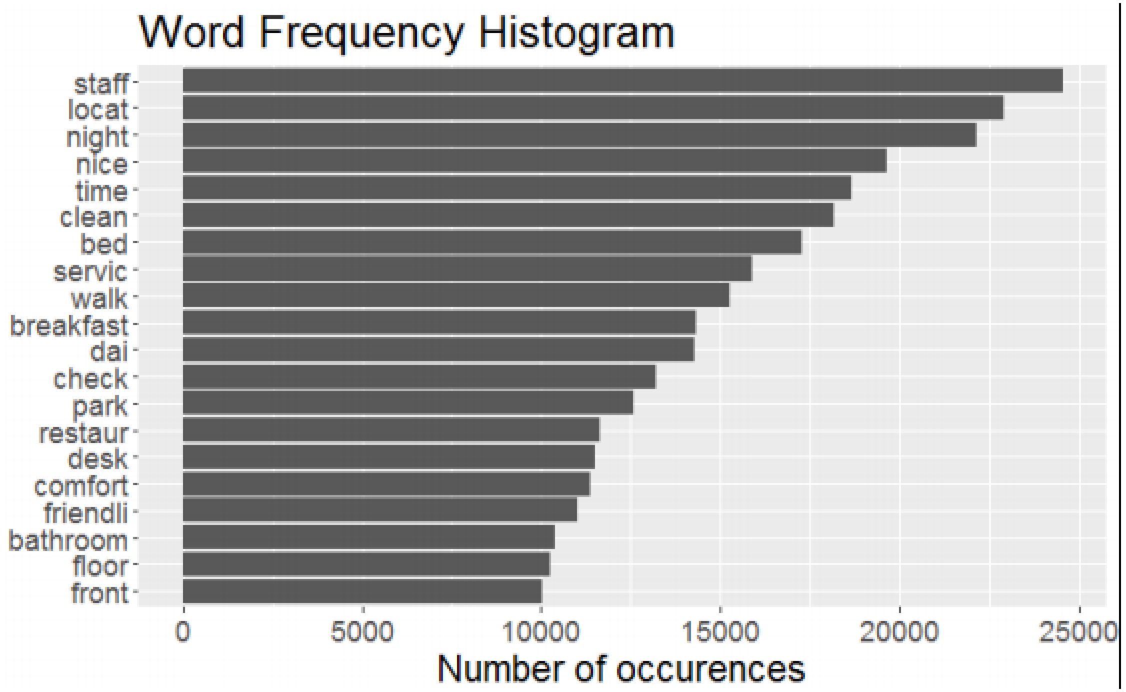
INFOX
educación sin fronteras

Características de la visualización

- Ejemplo, este es un gráfico simple en que sólo el tamaño de la palabra va a importar

Gráficos vs Palabras

- Cómo presentar la información?



70 palabras

20 palabras

- Algunos gráficos brindan más información que otros.

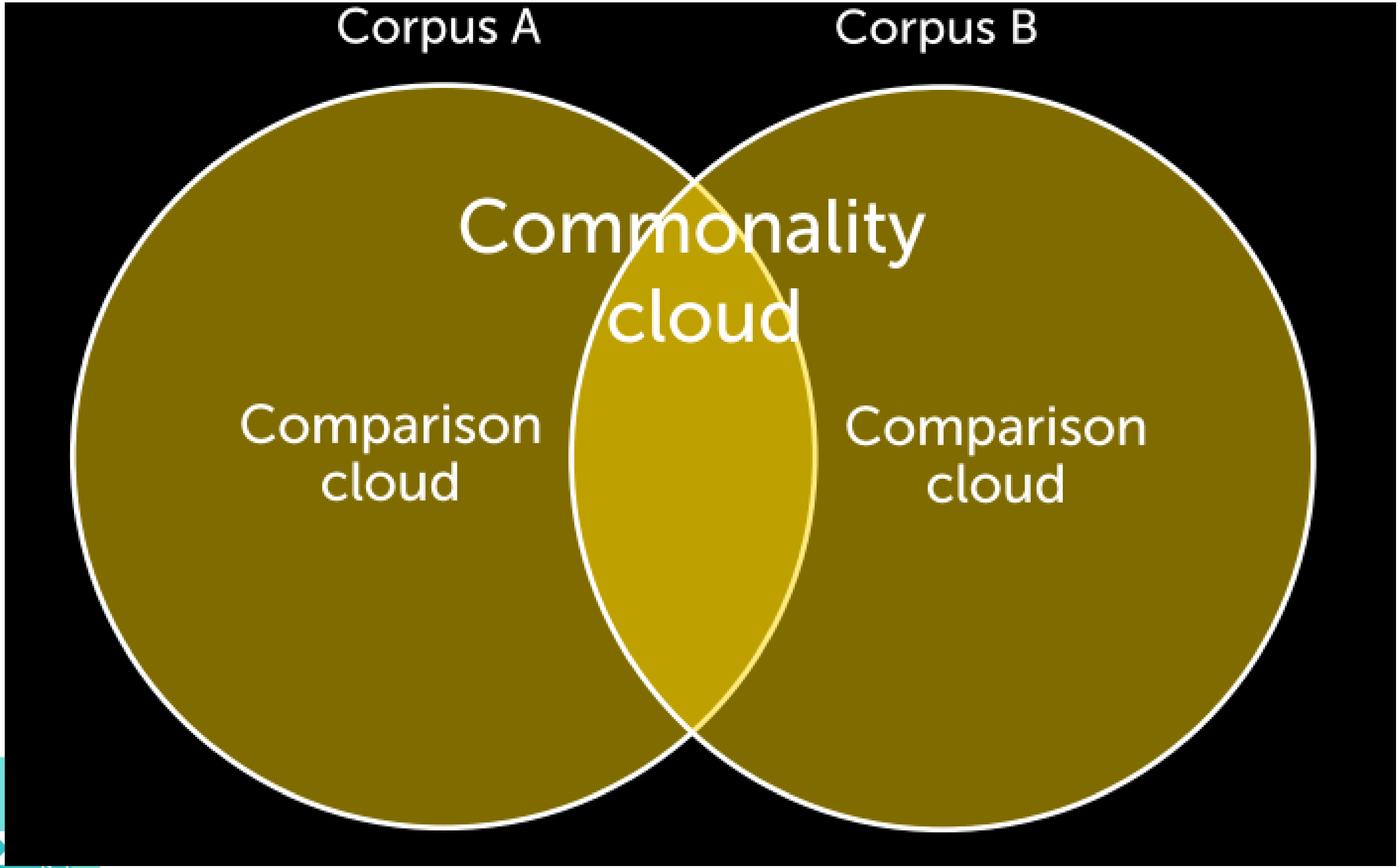
Contrasting grahps

Happy customers

Unhappy customers

Todavía hay muchas palabras que se entrelazan. Las personas usan mucho “location” y “staff”. Podemos resolver esto?

Wordcloud package



Qué palabras son las más usadas tanto en “happy” y “unhappy” reviews? Commonality cloud

INFOX
educación sin fronteras

Un gráfico de contraste: the comparison cloud

- “Staff” y “location” son usados en ambos, pero más en comentarios “happys” que “unhappys”

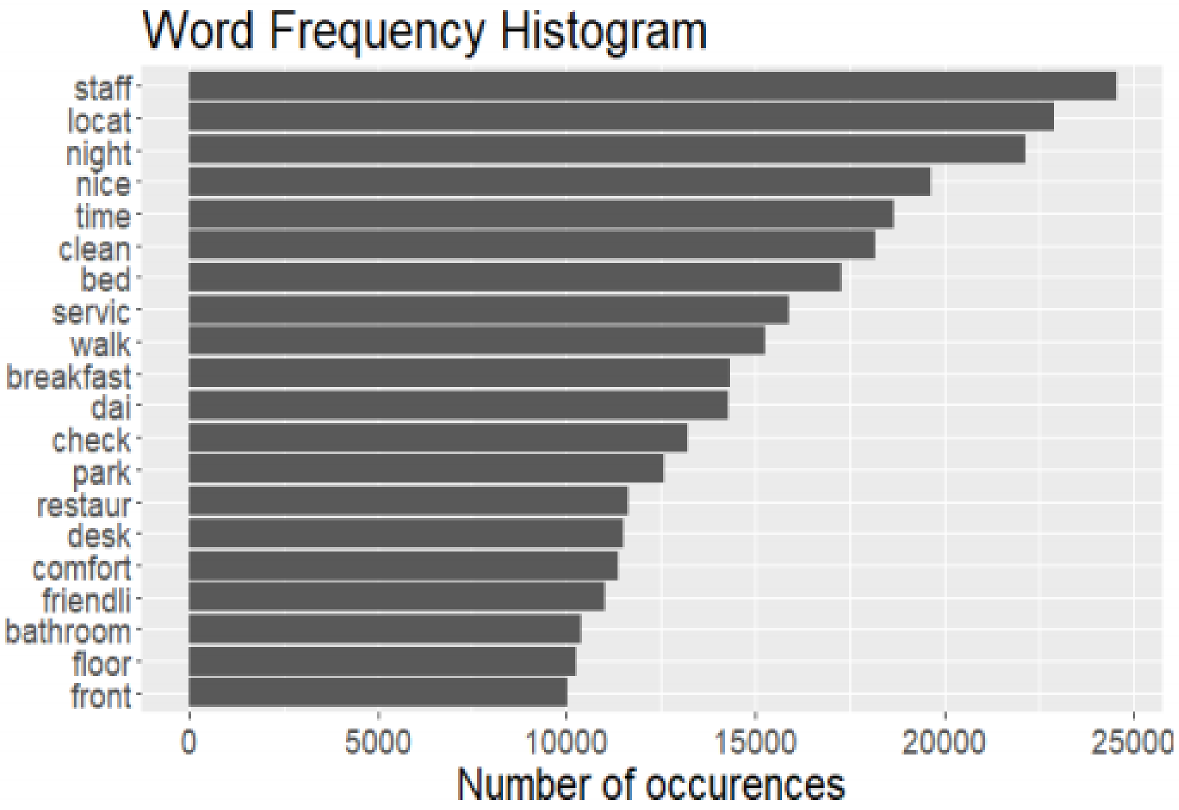
- La selección de palabras se usa en la frecuencia y no en el poder de discriminación

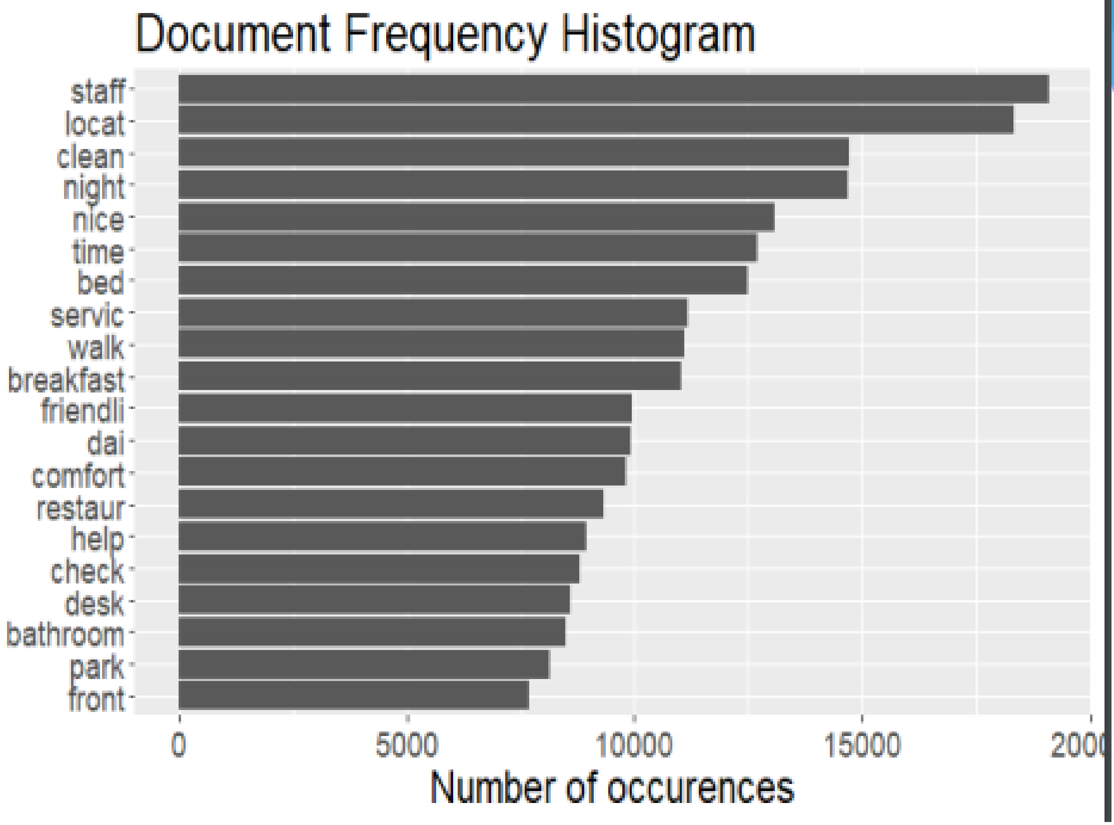
TF-IDF: which words differentiate

- TF-IDF = Term Frequency – Inverse Document Frequency
- Term Frequency: el número de veces que un término ocurre en todos los documentos. Cuando un documento varía mucho en la longitud, dividimos por el número total de palabras.
- Document Frequency: El número de documentos que contiene el término
- Qué palabras describen mejor el contenido de una muestra de reviews?
- Las palabras que ocurren más a menudo en esa muestra de reviews
- Corregido por qué tan frecuentemente aparece esta palabra
- Dividimos por cuántos documentos contiene esa palabra

INFOX
educación sin fronteras

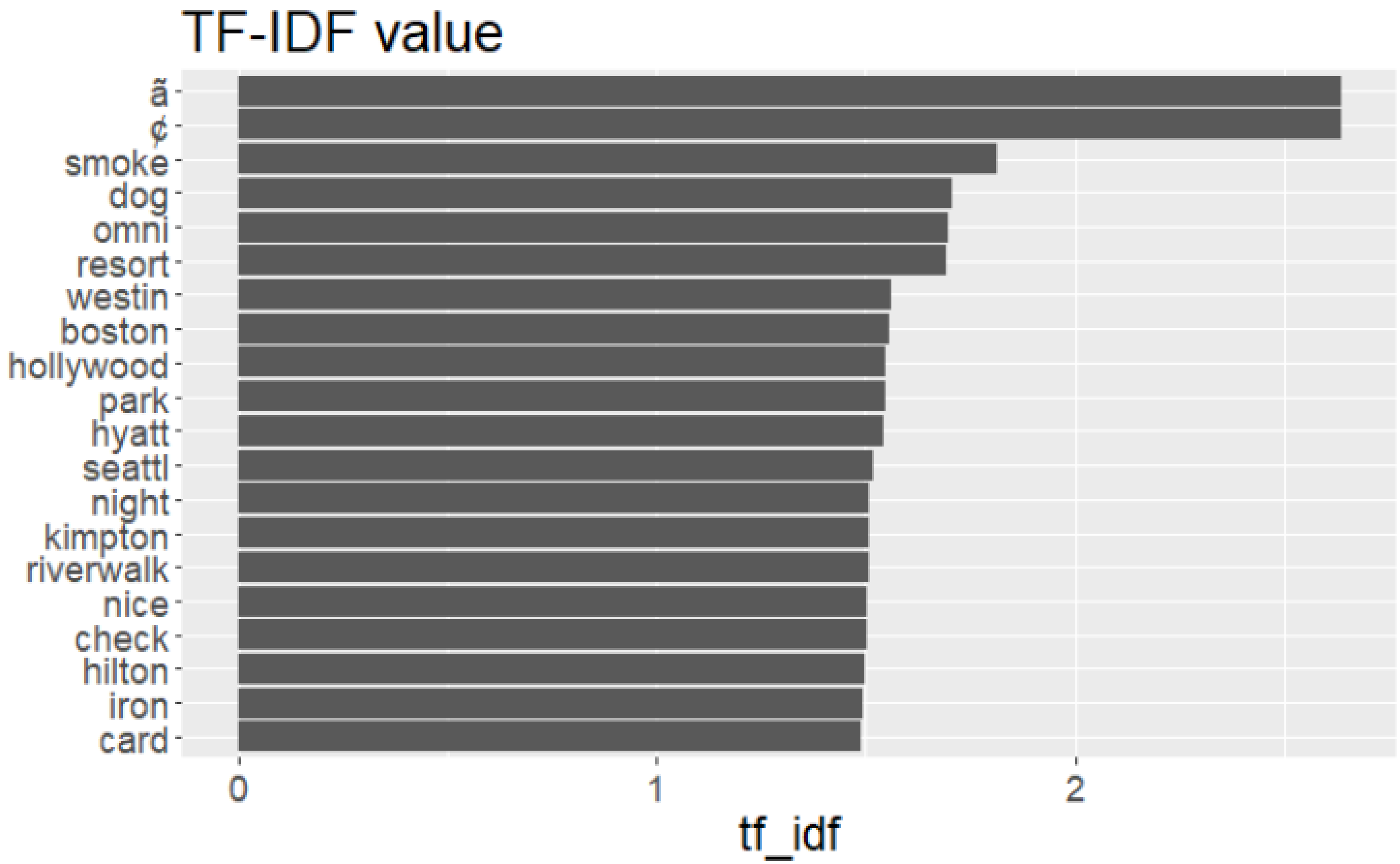
TF-IDF: which words differentiate





INFOX
educación sin fronteras

TF-IDF: which words differentiate



La frecuencia nos dice de qué se tratan los reviews.
Qué significa un alto valor de TF-IDF?

Nota: Una palabra no puede ocurrir si no está en un documento, por lo que siempre TF-IDF>1

Maximizando el contraste

- TF-IDF corrige por el total de ocurrencias de palabras en todos los reviews.

- Podemos generar un fuerte contraste en los reviews cuando comparamos diferentes muestras?

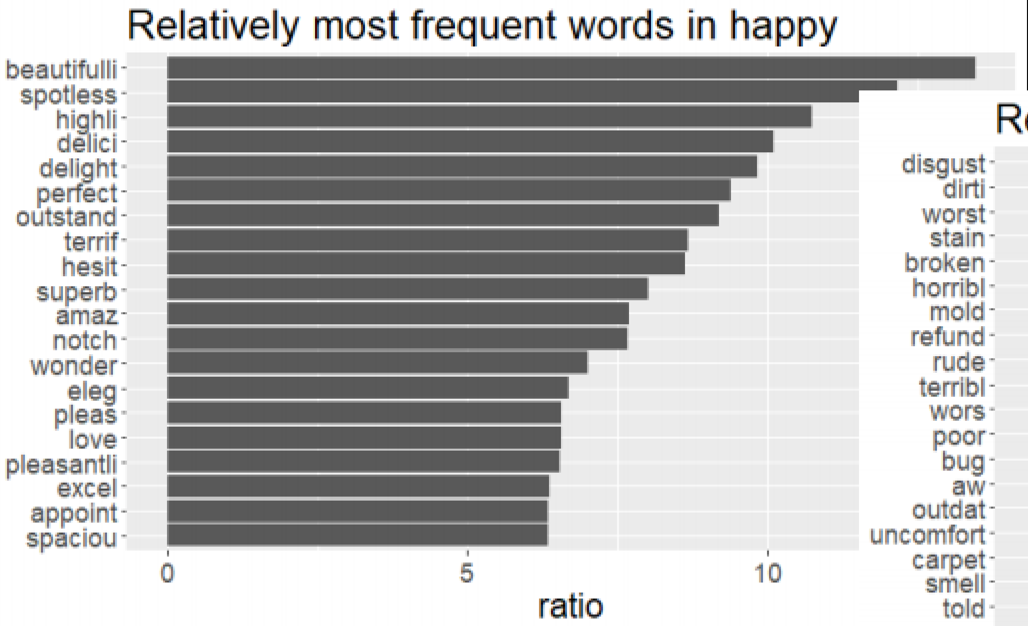
- Cuando contrastamos una muestra de reviews un benchmark directo nos dará más fuertes contrastes.

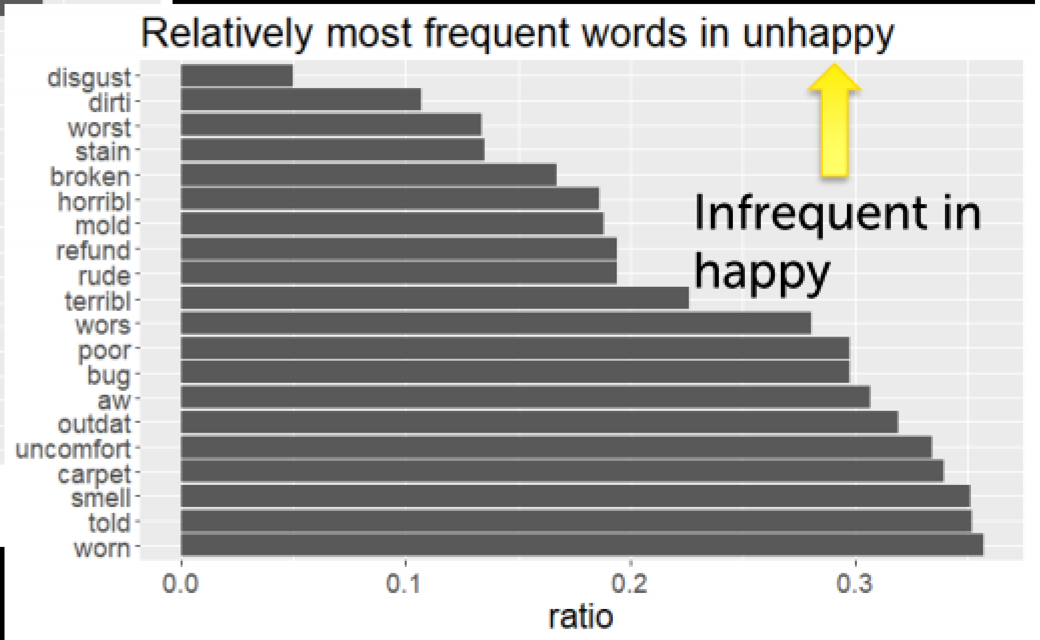
- Grupo A es caracterizado – relativo al grupo B – por palabras que ocurren mas frecuentemente en A relativo a B

- TF(A)/TF(B)
- También puede hacerse TF-IDF(A)/TF-IDF(B) ya que la parte de IDF se cancela
- Palabras que ocurren generalmente seguido obtienen una calificación

Maximizando el contraste

- Podemos contrastar un gráfico que mejor contraste “happy” y “unhappy”?
- Qué información debería representar?
- Frecuencia relativa de las palabras en “happy” relativa a los “unhappy” reviews?





Visualizando contraste en el texto

- Comparison cloud
- Mostrar que palabras ocurren más en cada muestra
- Mayormente muestra las palabas más frecuentes
- No necesariamente las palabras más informativas/ predictivas

- Frecuencias relativas
- Resaltar las palabras más…
- Informativos
- Menos comúnes

- Se pueden generar visualizaciones bonitas.

INFOX
educación sin fronteras

Multidimensional scaling (MDS)

¿Podemos hacer la locación más informativa?

- Palabras que están cerca también son similares
- Se requiere información para palabras similares
- Para pasar de la data a la información se requiere análisis
- Hay 2 métodos

1. Factor Analysis
2. Multidimensional scaling

¿Qué significa la distancia en un mapa?

INFO
educación sin fronteras

¿Qué significa la distancia en un mapa?

	Anchorage	Atlanta	Baltimore	Boston	Chicago	Houston	Las Vegas	Los Angeles
Anchorage	0	5471.52	5392.82	5416.45	4584.33	5260.73	3690.71	3763.14
Atlanta	5471.52	0	927.35	1505.11	944.4	1126.72	2801.21	3108.01
Baltimore	5392.82	927.35	0	577.85	973.23	2010.47	3377.44	3722.45
Boston	5416.45	1505.11	577.85	0	1366.63	2578.59	3809.81	4166.43
Chicago	4584.33	944.4	973.23	1366.63	0	1510.62	2443.78	2799.8
Houston	5260.73	1126.72	2010.47	2578.59	1510.62	0	1971.57	2205.12
Las Vegas	3690.71	2801.21	3377.44	3809.81	2443.78	1971.57	0	367.91
Los Angeles	3763.14	3108.01	3722.45	4166.43	2799.8	2205.12	367.91	0

¿Podemos hacer algo relacionado con palabras?

- Qué podría medir si las palabras son cercanas en un Word vector model?

- Podemos medir si las palabras pertenecen a un mismo vecindario?

- Otras ideas de cómo medir qué tan cercanas son las palabras?

- Qué data necesitaríamos?

Word similarity

- Las palabras son más similares o más relacionadas cuando …
- Ocurren juntas en un review
- Ocurren juntas en una oración
- Están juntas la una de la otra en un texto

INFOX
educación sin fronteras

Word similarity

- Qué tan a menudo ocurren las palabras juntas en un review?
- Qué matriz/ tabla captura esto?
- The Burt table
- Y esta tabla cómo se parece?

	locat	night	nice	clean	bed	servic	walk	Break fast	check
locat	18327	7247	6646	7668	6427	5380	6696	5546	4093
night	7247	14688	5604	6092	6001	4399	4973	4759	4464
nice	6646	5604	13067	5690	5298	3885	4421	4322	3531
clean	7668	6092	5690	14724	5649	3894	4908	5076	3586
bed	6427	6001	5298	5649	12505	3729	4362	4031	3833
servic	5380	4399	3885	3894	3729	11181	3128	3251	3148
walk	6696	4973	4421	4908	4362	3128	11109	3740	2960
breakfast	5546	4759	4322	5076	4031	3251	3740	11018	2715
check	4093	4464	3531	3586	3833	3148	2960	2715	8808
park	4385	3900	3405	3499	3116	2405	3419	2838	2335
restaur	5211	3974	3897	3812	3541	3352	3976	3313	2342

INFOX
educación sin fronteras

Word similarity

- Cómo visualizamos esto?
- Palabras que ocurren juntas usualmente están cerca
- MDS busca poner las palabras en posiciones que reflejan la distancia entre ellas

Objetivo

Minimize sum over all pairs of $(\text{distance in map} - \text{distance in data})^2$

We now have similarities

- Needs to be transformed to distances

Smacof package

Transform similarity values to distances

Sim2diss co-occurrence option

S is similarity matrix
Rowsum, colsum and totalsum are corresponding sums

Normalize similarities

- S <- (totalsum * S) / (rowsum %*% t(colsum))

Transform similarities to distances

- dissmat <- 1 / (1 + S)

Open question:
Should diagonals in S contain 0 or document frequency value?

MDS

	Anchorage	Atlanta	Baltimore	Boston	Chicago	Houston	Las Vegas	Los Angeles
Anchorage	0	5471.52	5392.82	5416.45	4584.33	5260.73	3690.71	3763.14
Atlanta	5471.52	0	927.35	1505.11	944.4	1126.72	2801.21	3108.01
Baltimore	5392.82	927.35	0	577.85	973.23	2010.47	3377.44	3722.45
Boston	5416.45	1505.11	577.85	0	1366.63	2578.59	3809.81	4166.43
Chicago	4584.33	944.4	973.23	1366.63	0	1510.62	2443.78	2799.8
Houston	5260.73	1126.72	2010.47	2578.59				
Las Vegas	3690.71	2801.21	3377.44	3809.81				
Los Angeles	3763.14	3108.01	3722.45	4166.43				

Las distancias obtenidas de las co-ocurrences

```
distances <- sim2diss(Burt_fcm, method = "cooccurrence")
```

	locat	night	nice	clean	bed
locat	0.309579	0.504355	0.485693	0.465003	0.503885
night	0.504355	0.310603	0.501253	0.495396	0.493961
nice	0.485693	0.501253	0.268385	0.472182	0.484806
clean	0.465003	0.495396	0.472182	0.268531	0.483786
bed	0.503885	0.493961	0.484806	0.483786	0.293112
servic	0.4965	0.519748	0.510511	0.52493	0.53054

Differences look small, but are informative

INFOX
educación sin fronteras

Word similarity

- ¿Qué tan seguido las palabras ocurren juntas en un review?

Note the large number of words being present here!

Word similarity based on distance

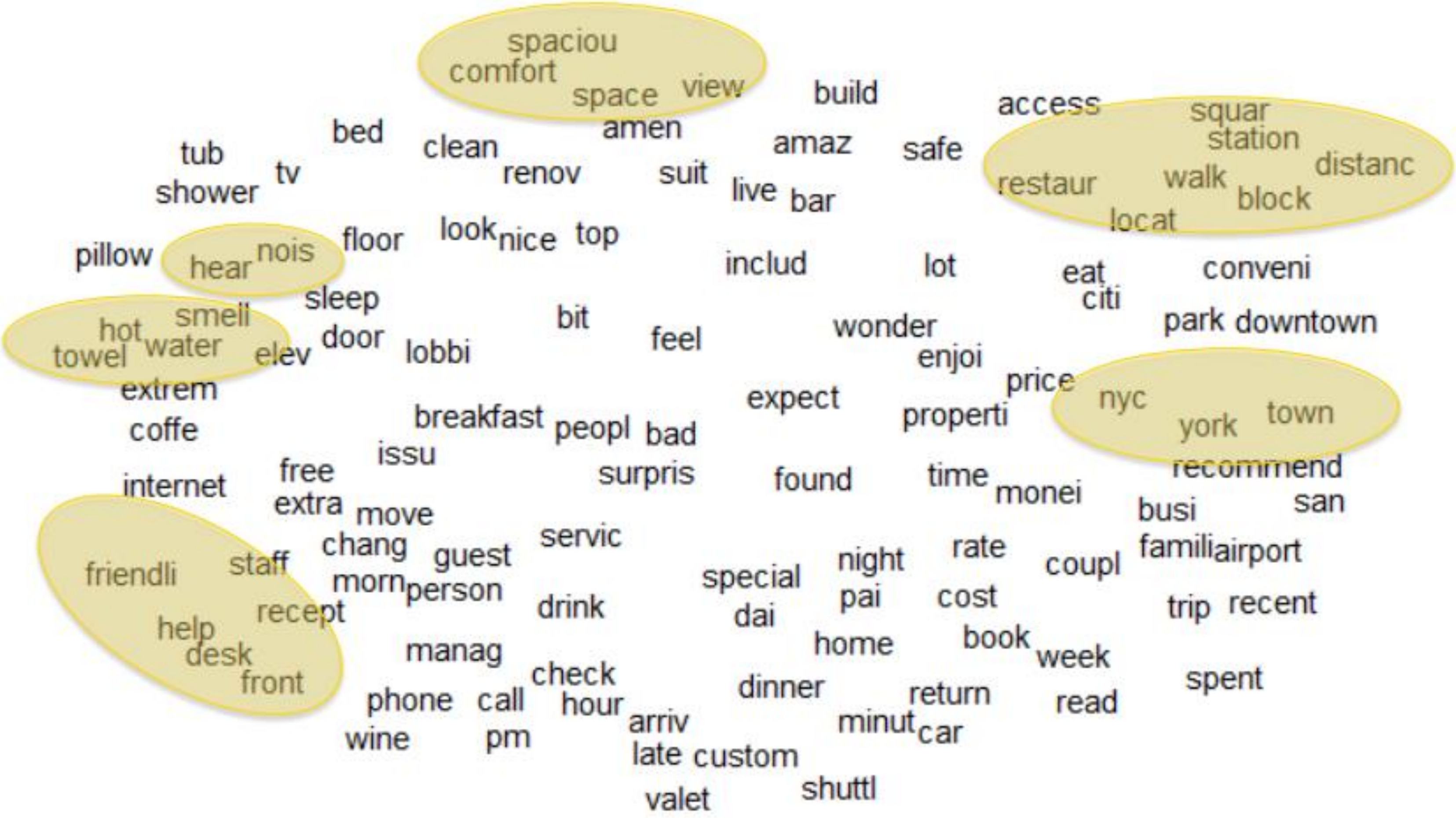
- Las palabras están relacionadas cuando están cercas las unas de las otras

- Count how often they are within K words
  - N-gram
  - Before or after data cleaning

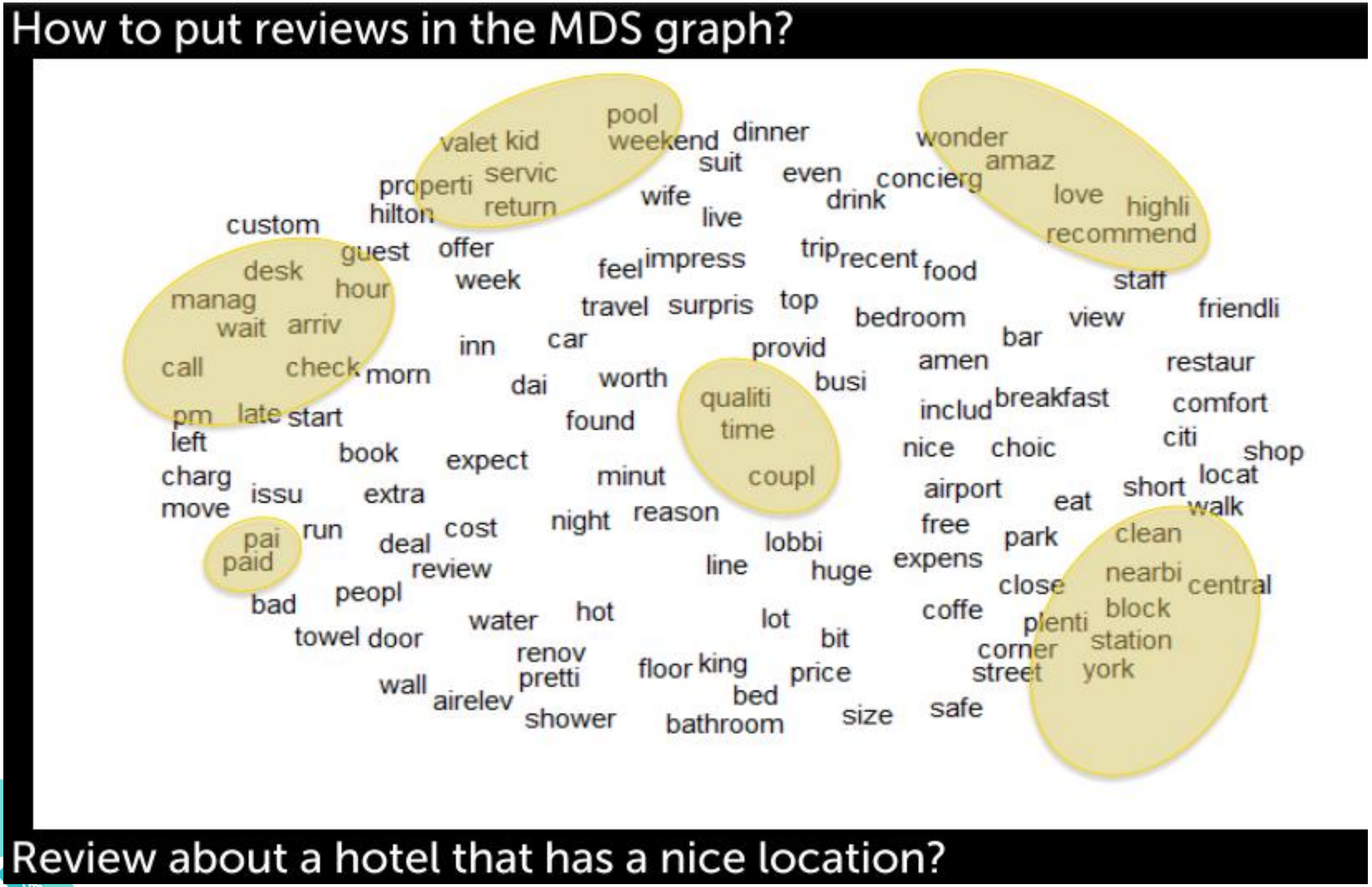
Neighbors in cleaned data: what will happen?

	night	nice	time	clean	bed	servic	walk	breakf	check	park	desk
night		605	844	670	738	582	559	360	738	1107	293
nice	605		585	2224	1379	729	551	1008	523	415	584
time	844	585		570	323	608	984	380	1030	439	335
clean	670	2224	570		2216	804	322	616	396	229	309
bed	738	1379	323	2216		333	254	301	278	112	409
servic	582	729	608	804	333		179	763	493	251	551
walk	559	551	984	322	254	179		268	232	827	171
breakfast	360	1008	380	616	301	763	268		220	299	156
check	738	523	1030	396	278	493	232	220		311	927
park	1107	415	439	229	112	251	827	299	311		126
desk	293	584	335	309	409	551	171	156	927	126	

MDS para palabras en vecindarios



Poniendo reviews en un gráfico de MDS



¡GRACIAS!

INFOX
educación sin fronteras