FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Grado en Inteligencia Artificial

Textual Analysis – CA Approach

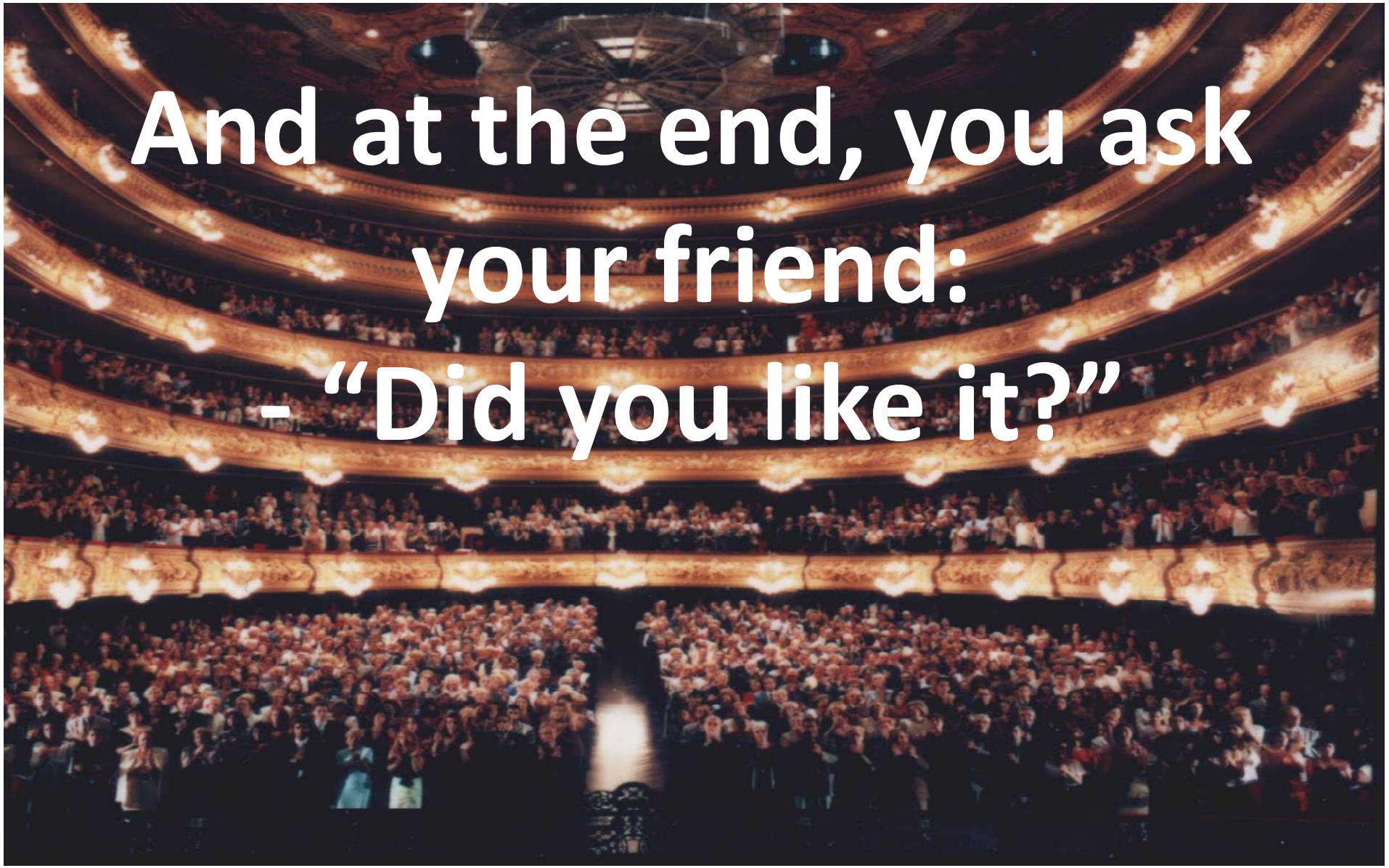
Prof. Dante Conti

Prof. Sergi Ramírez

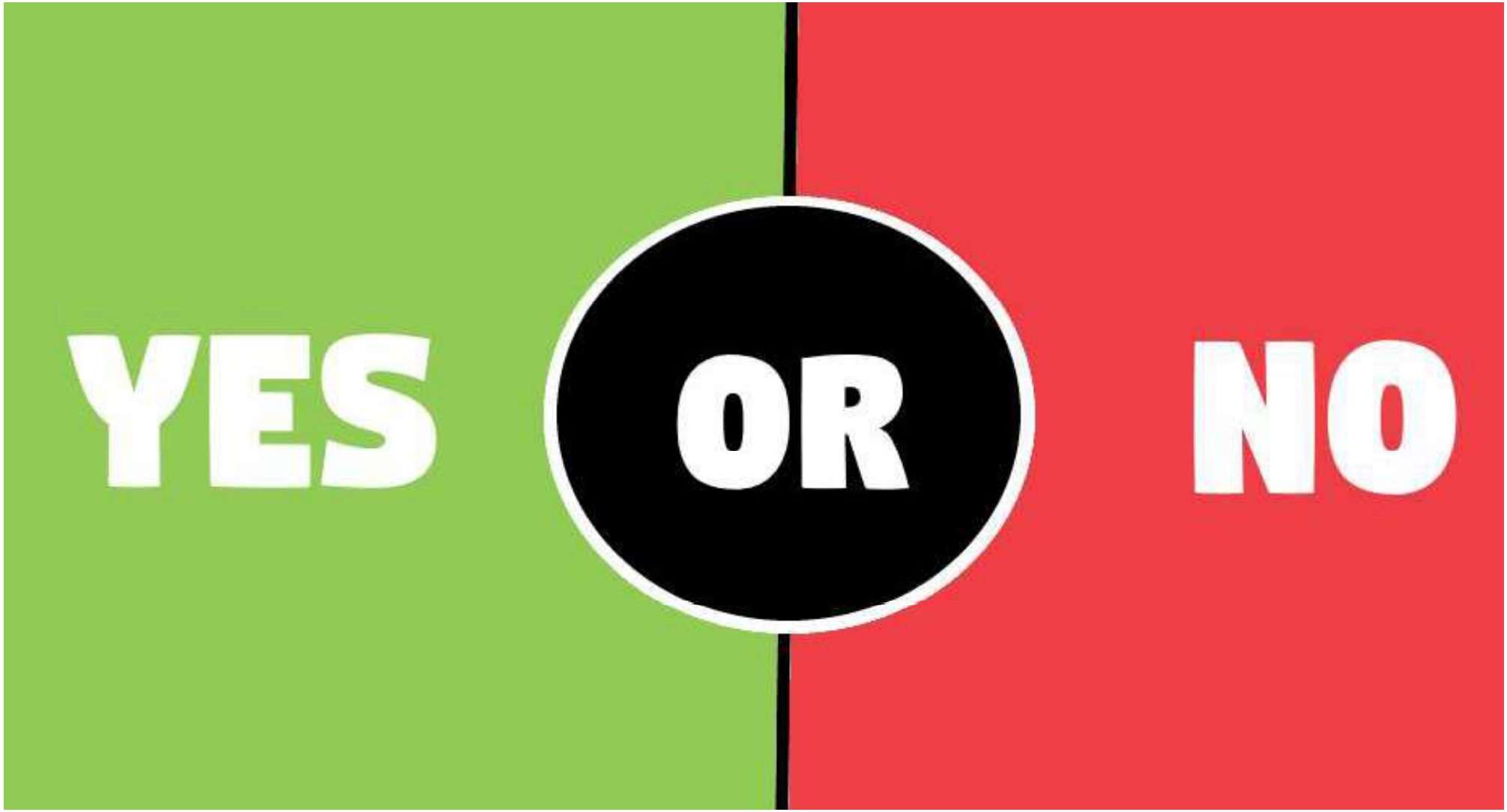
Outline

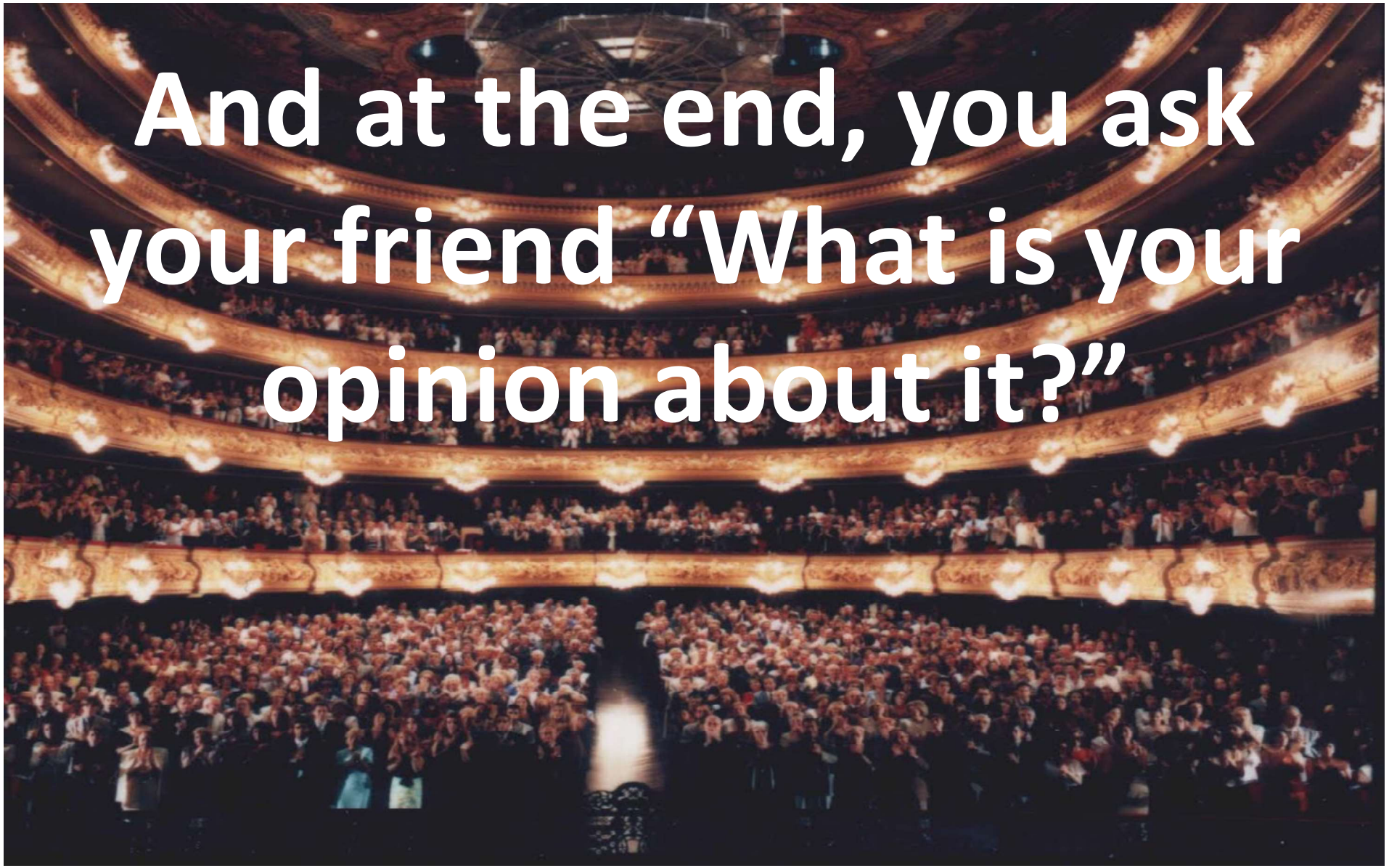
1. Introduction
2. Methods and approaches
3. Correspondence Analysis applied to textual data
4. Example 1: Definition of health
5. Example 2: Comparison of factor analysis methods
6. Summary and conclusions

Introduction


And at the end, you ask your friend:
- “Did you like it?”

Introduction



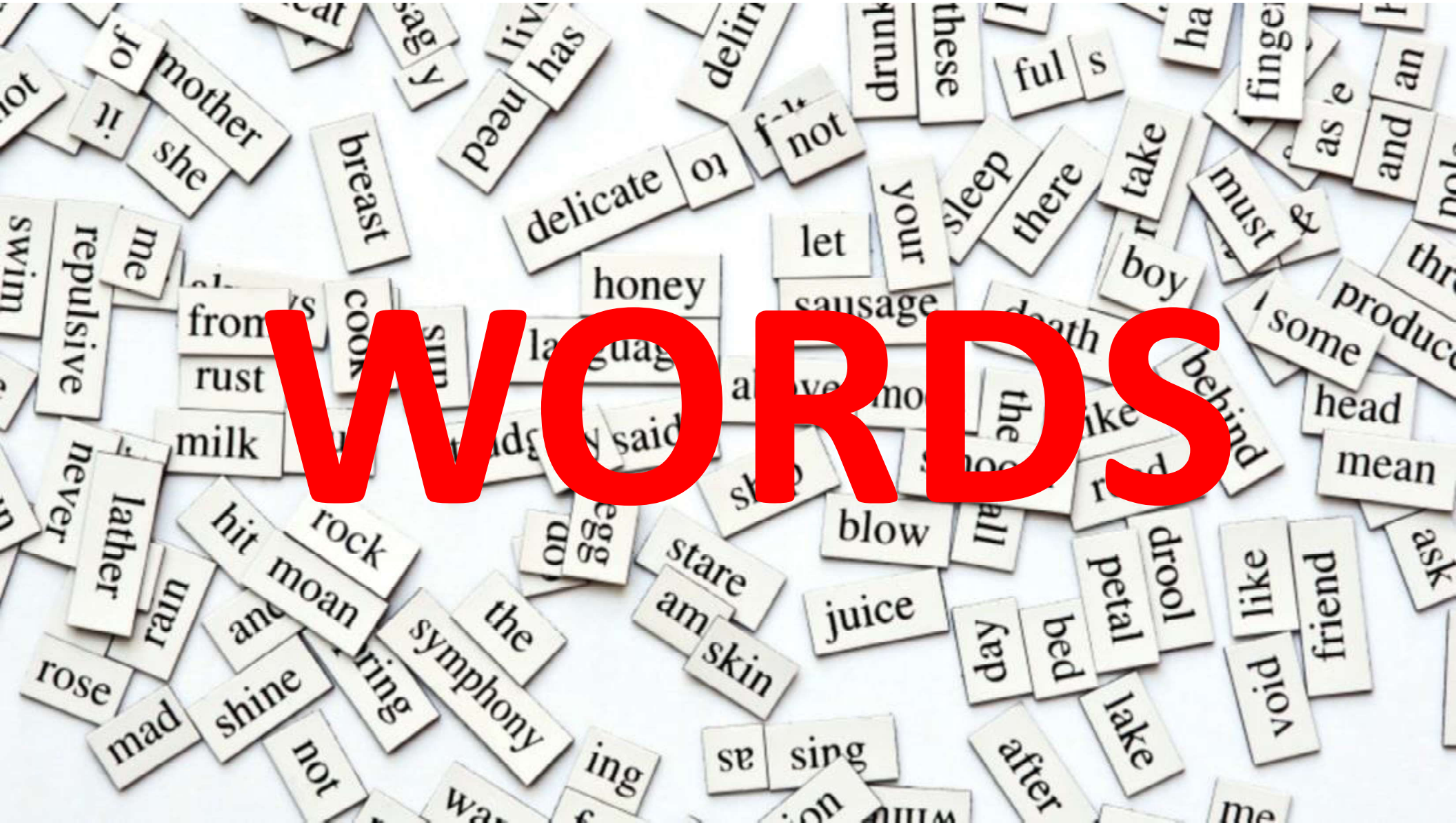

And at the end, you ask your friend “What is your opinion about it?”

Introduction

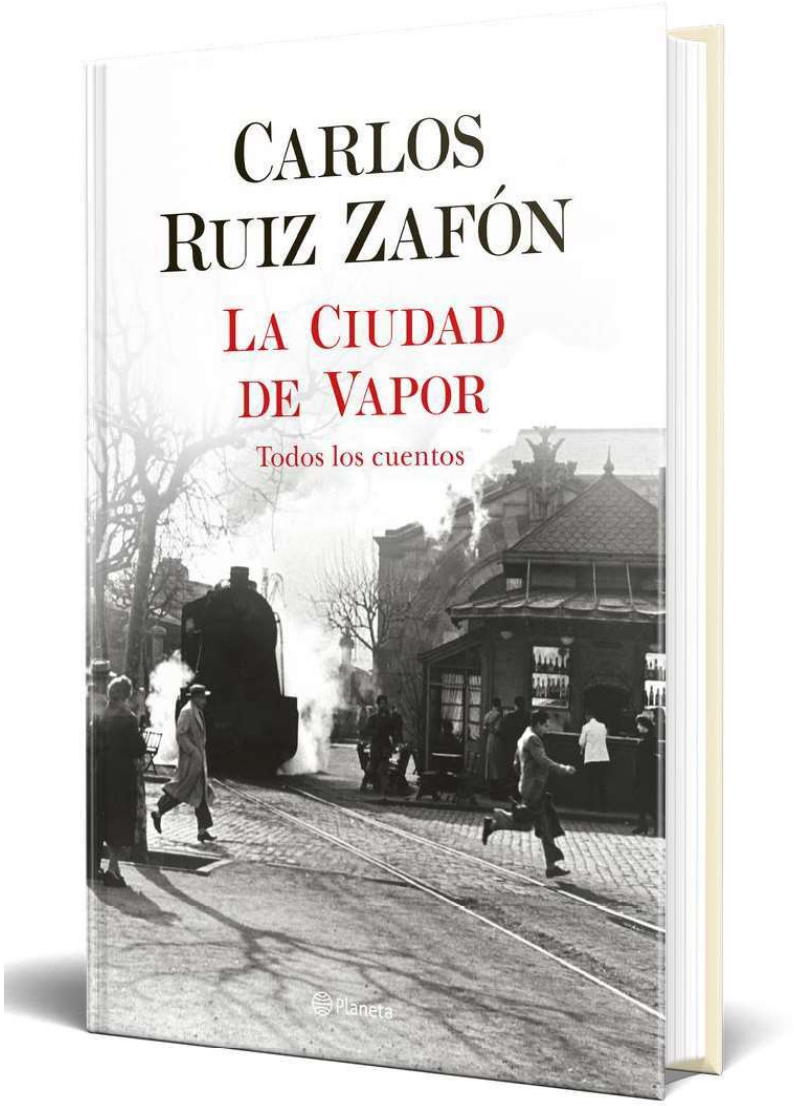
What do you expect when you go on holidays?

What do you expect when you go on holidays?

Words



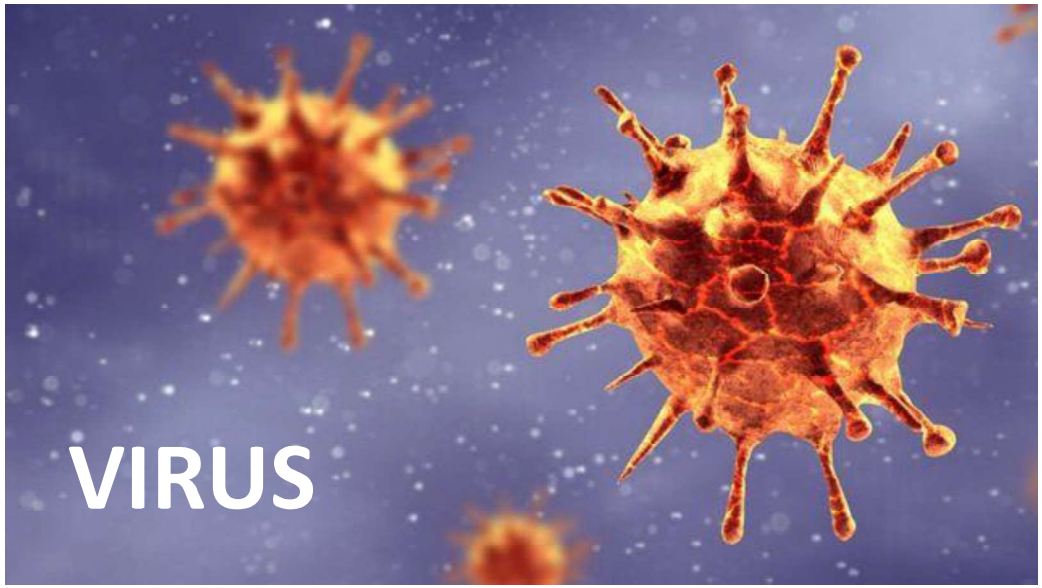
Books



«Puedo conjurar rostros de chiquillos del barrio de la Ribera con los que a veces jugaba o peleaba en la calle, pero ninguno que quisiera rescatar del país de la indiferencia. Ninguno excepto el de Blanca.»

Open-ended questions

What comes to your mind when you think about “COVID-19”?



Speeches

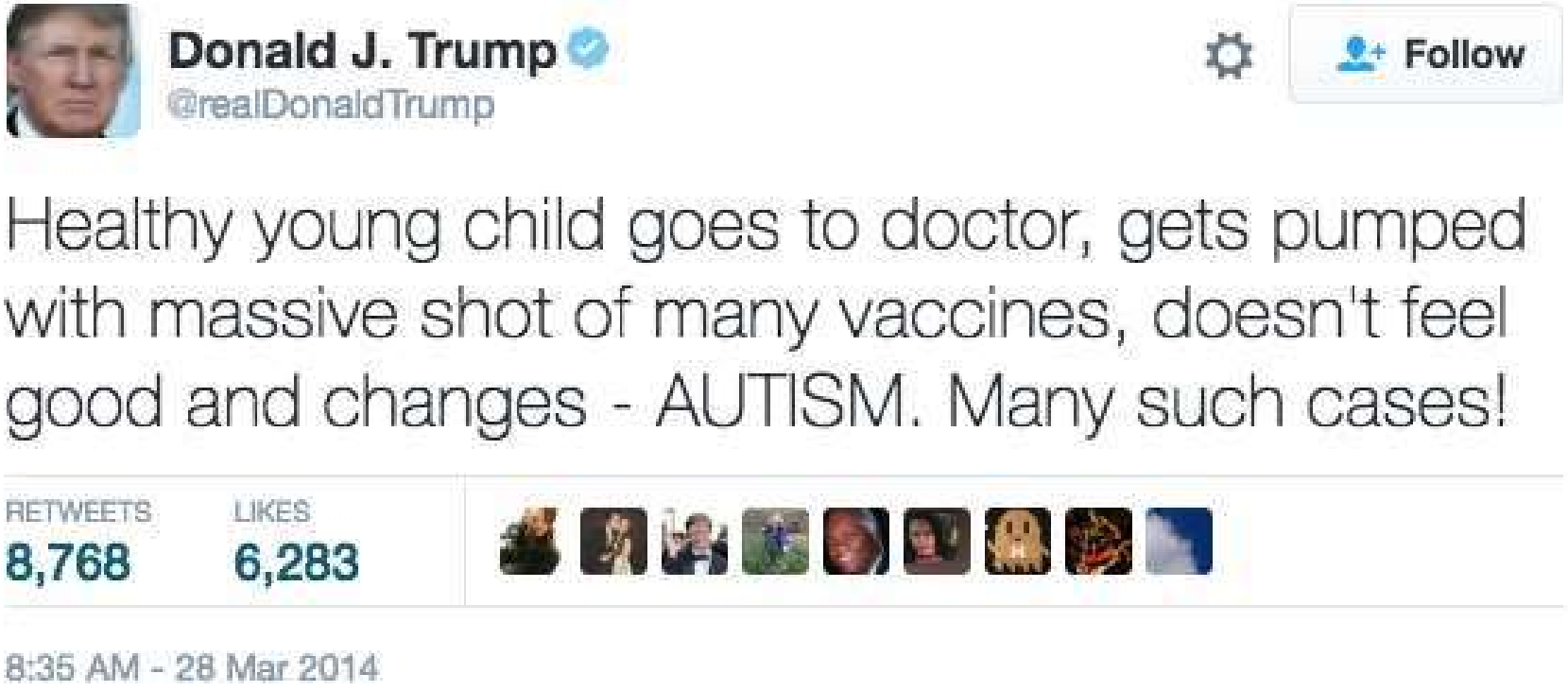
I Have a Dream (delivered 28 August 1963, at the Lincoln Memorial, Washington D.C.)

I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation.

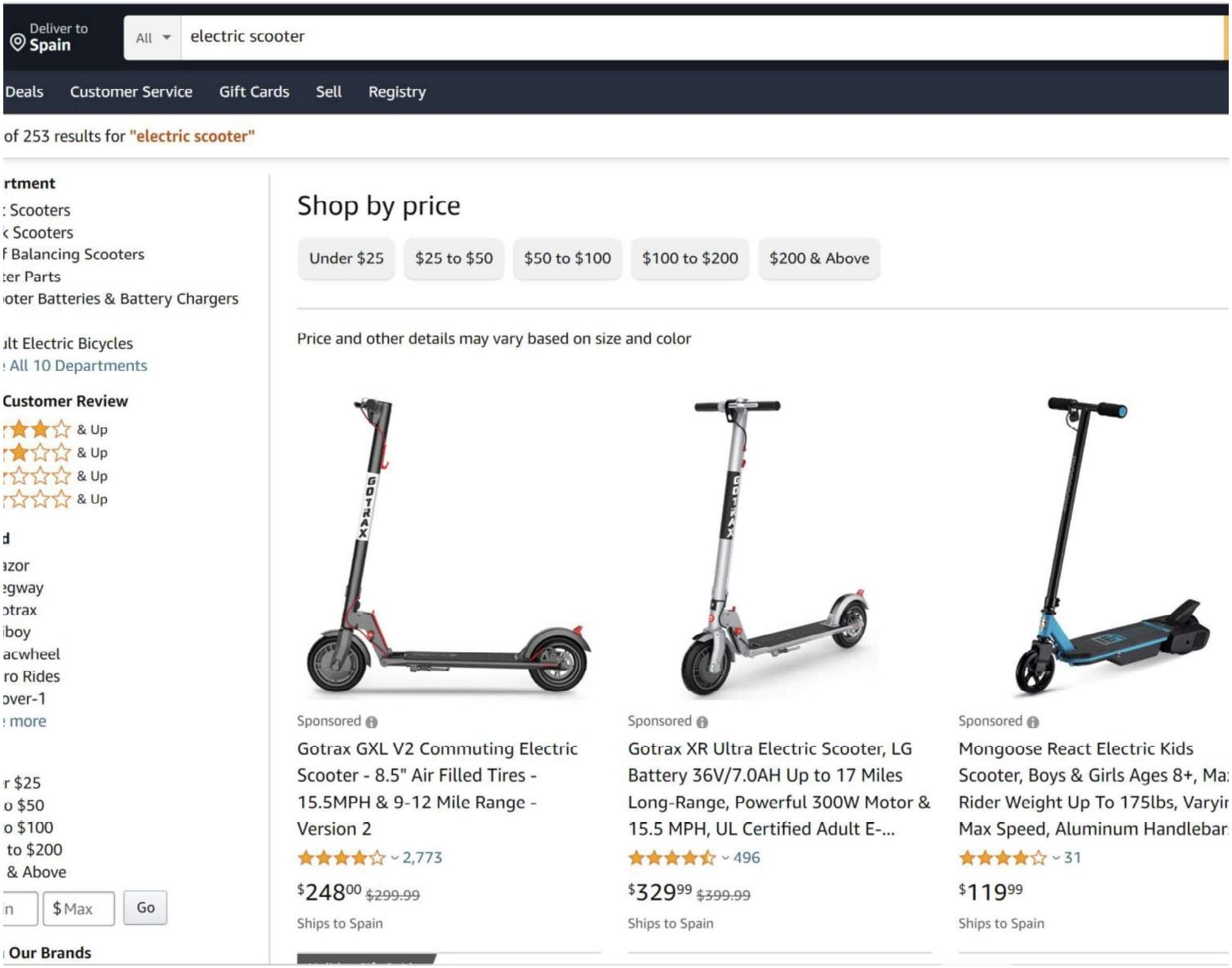
Five score years ago, a great American, in whose symbolic shadow we stand today, signed the Emancipation Proclamation. This momentous decree came as a great beacon light of hope to millions of Negro slaves who had been seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of their captivity.

But one hundred years later, the Negro still is not free…….

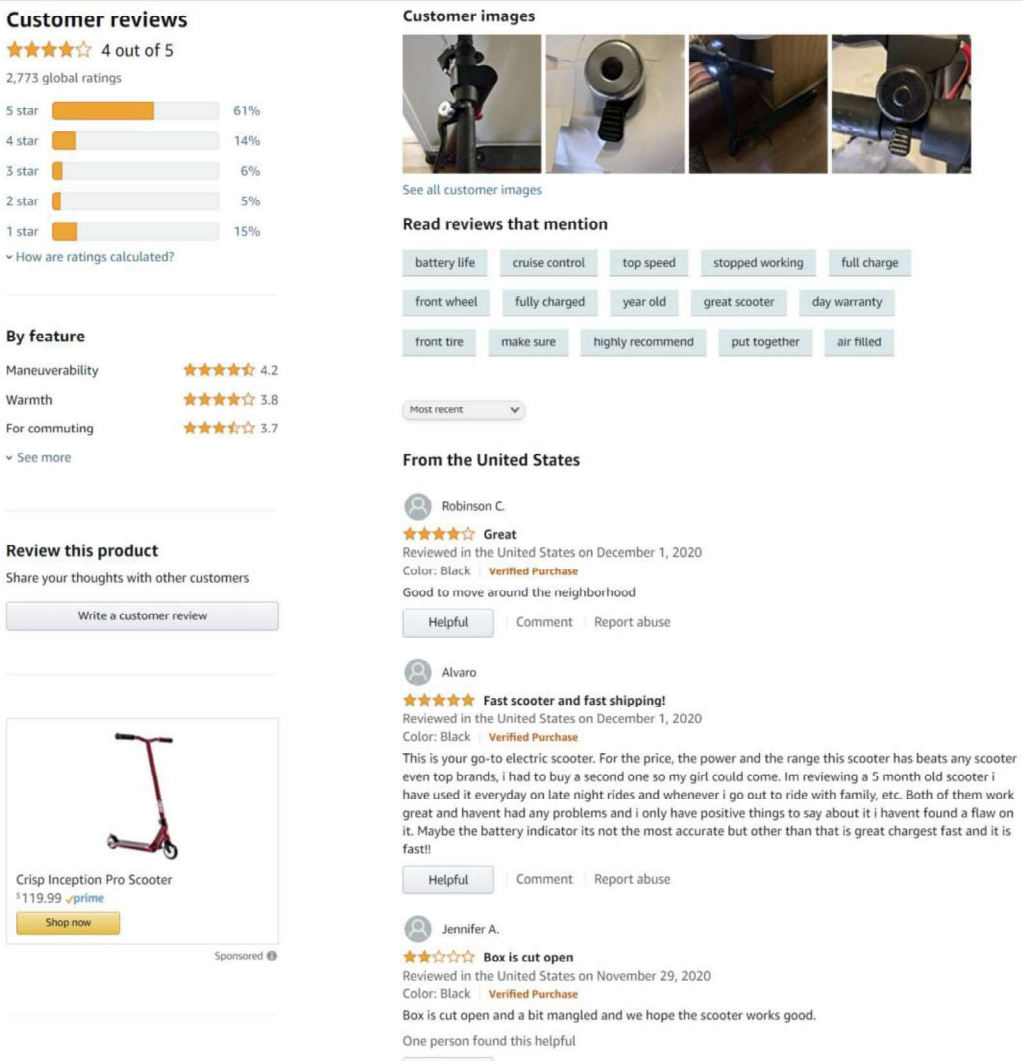
Social networks



I would like to buy a new electric scooter…



Customer reviews



Customer reviews

4 out of 5

2,773 global ratings

5 star 61%
4 star 14%
3 star 6%
2 star 5%
1 star 15%

How are ratings calculated?

By feature

Maneuverability 4.2
Warmth 3.8
For commuting 3.7

See more

Review this product

Share your thoughts with other customers

Write a customer review

Crisp Inception Pro Scooter
$119.99 prime
Shop now

Sponsored

Customer images

See all customer images

Read reviews that mention

battery life | cruise control | top speed | stopped working | full charge
front wheel | fully charged | year old | great scooter | day warranty
front tire | make sure | highly recommend | put together | air filled

Most recent

From the United States

Robinson C.
Great
Reviewed in the United States on December 1, 2020
Color: Black | Verified Purchase
Good to move around the neighborhood
Helpful | Comment | Report abuse

Alvaro
Fast scooter and fast shipping!
Reviewed in the United States on December 1, 2020
Color: Black | Verified Purchase
This is your go-to electric scooter. For the price, the power and the range this scooter has beats any scooter even top brands, i had to buy a second one so my girl could come. Im reviewing a 5 month old scooter i have used it everyday on late night rides and whenever i go out to ride with family, etc. Both of them work great and havent had any problems and i only have positive things to say about it i havent found a flaw on it. Maybe the battery indicator its not the most accurate but other than that is great chargest fast and it is fast!!
Helpful | Comment | Report abuse

Jennifer A.
Box is cut open
Reviewed in the United States on November 29, 2020
Color: Black | Verified Purchase
Box is cut open and a bit mangled and we hope the scooter works good.
One person found this helpful

Textual data: objectives

- To bring up the subject
- To determine the structure of the text: localizing the ruptures
- To visualize the proximity between documents and words
- To index the documents
- To construct a database with documents for a subsequent automatic interrogation

Textual data: methods

- Classical methods: frequency of words
- Multidimensional methods: Factor analysis
- Recent methods: Machine learning
- Two important points: codification and distance between the elements

Example 1: Definition of health

- Surveys often include open-ended questions as a complement to classical closed questions to address complex and little known topics.
- Example: survey intended to better know the definitions of health that the non-experts give
  - 392 respondents who answered through free-text comments
  - The respondents' characteristics: age in groups (under 21, 21–35, 36–50 and over 50), gender (man and woman) and health condition (poor, fair, good and very good health)

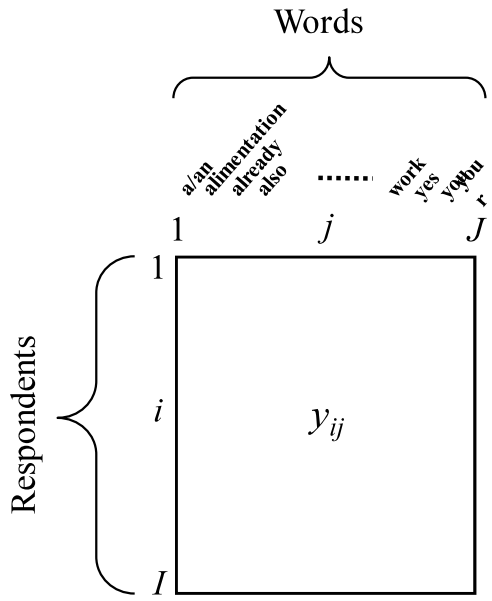
Some data pre-treatment steps

To obtain the final word frequency table to analyze, we :

- Replace special characters from the text (for example, replacing "/", "@" and "|" with space).
- Remove unnecessary white space
- Convert the text to lower case
- Convert the verbs to infinitive form
- Remove common stopwords like "the", "we", etc.

Codification: lexical table

- The (I×J) matrix **Y**, with generic term y_{ij}, contains the frequency of the J words in the I respondents' answers.
- Stopwords and lemmatization
- The minimum threshold for words: 10
- 115 different words and 7751 occurrences (N)

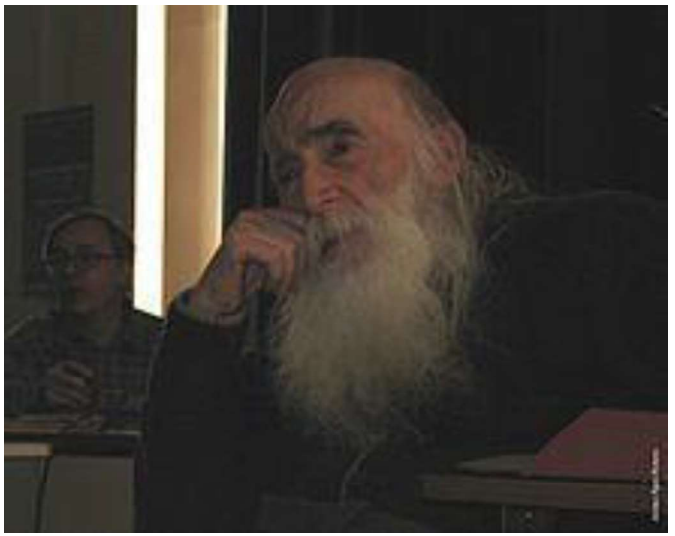


Analysis of lexical table

- Correspondence Analysis (CA) is the reference factorial method
  - Optimal visualization of the similarities among the free answers
  - Optimal visualization of the similarities among the words
  - The associations between words and free answers.

CA applied to textual data

- Correspondence analysis (CA) was proposed by *Benzecri (1973, 1981)* to deal with textual data.



- The term "correspondence analysis" stems from the fact that lexical table links two corresponding sets:
  - one represented by the rows, and
  - other represented by the columns, playing symmetric roles.

CA steps: Step 1

- The (I×J) frequency table **Y** is transformed into a proportion table

$$\mathbf{P} = \mathbf{Y}/N$$

with generic term $p_{ij} = y_{ij}/N$.

- The (I×I) diagonal matrix **D** stores the row margins $p_{i\bullet} = \Sigma_{j \in J} p_{ij}$
- The (J×J) diagonal matrix **M** stores the column margins $p_{\bullet j} = \Sigma_{i \in I} p_{ij}$

CA steps: Step 2

- Classical CA results can be obtained through a factor analysis applied to the (I×J) matrix

$$\mathbf{X} = \mathbf{D}^{-1}\mathbf{P}\mathbf{M}^{-1} = [q_{ij}] = [p_{ij}/p_{i\bullet}p_{\bullet j}]$$

with metrics/weights **M** and **D** (*Escofier and Pages, 1988).*

CA steps: Step 3

- The inertia axes in the row space satisfy the equation **X'DXMU = UΛ**
- The vector of coordinates of the rows **F = XMU**
- The inertia axes in the column space satisfy the equation **XMX'DV = VΛ**
- The vector of coordinates of the columns **G = X'DV**
- The simultaneous representation of rows and columns relies on the transition relationships

$$\mathbf{F} = \mathbf{V}\Lambda^{1/2}$$

$$\mathbf{G} = \mathbf{U}\Lambda^{1/2}$$

Interpretation rules

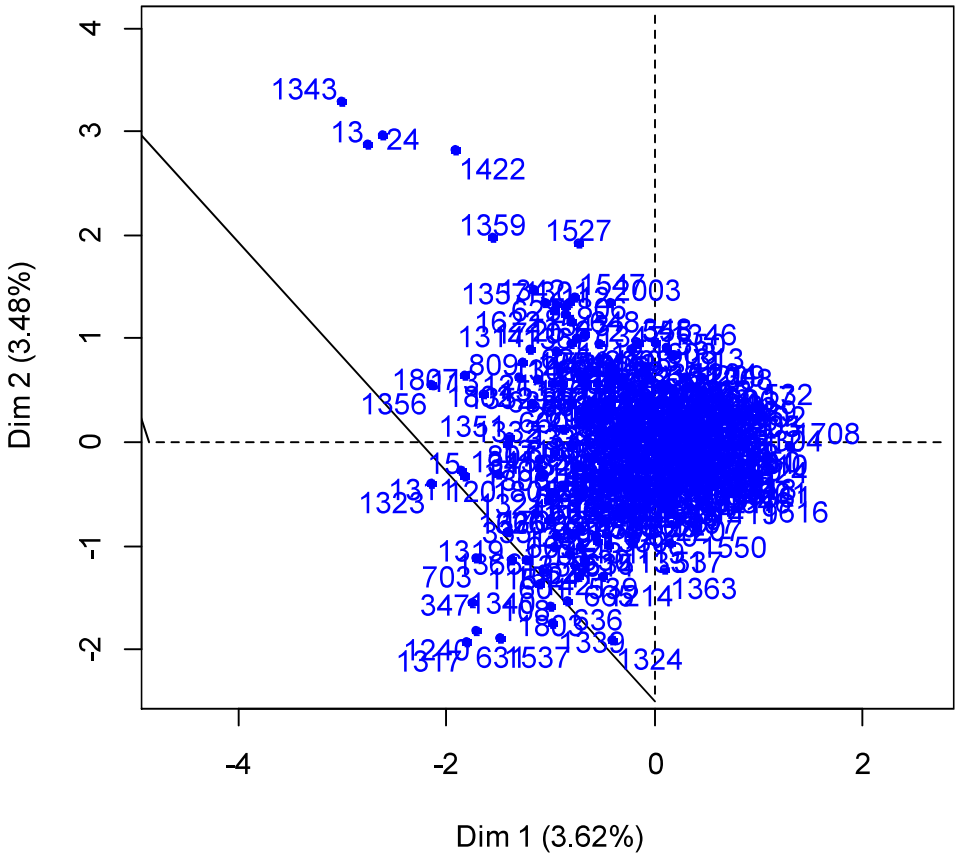
- For dimension *s* of this superimposed representation, up to the multiplicative factor $1/\sqrt{\lambda_s}$, a row i (resp. a column j) is at the barycentre of the columns (resp. of the rows), each column j (resp. each row i) having a weight $p_{ij}/p_{i\bullet}$ (resp. $p_{ij}/p_{\bullet j}$). This property is used to interpret the position of one row in relation to all of the columns and the position of one column in relation to all of the rows.

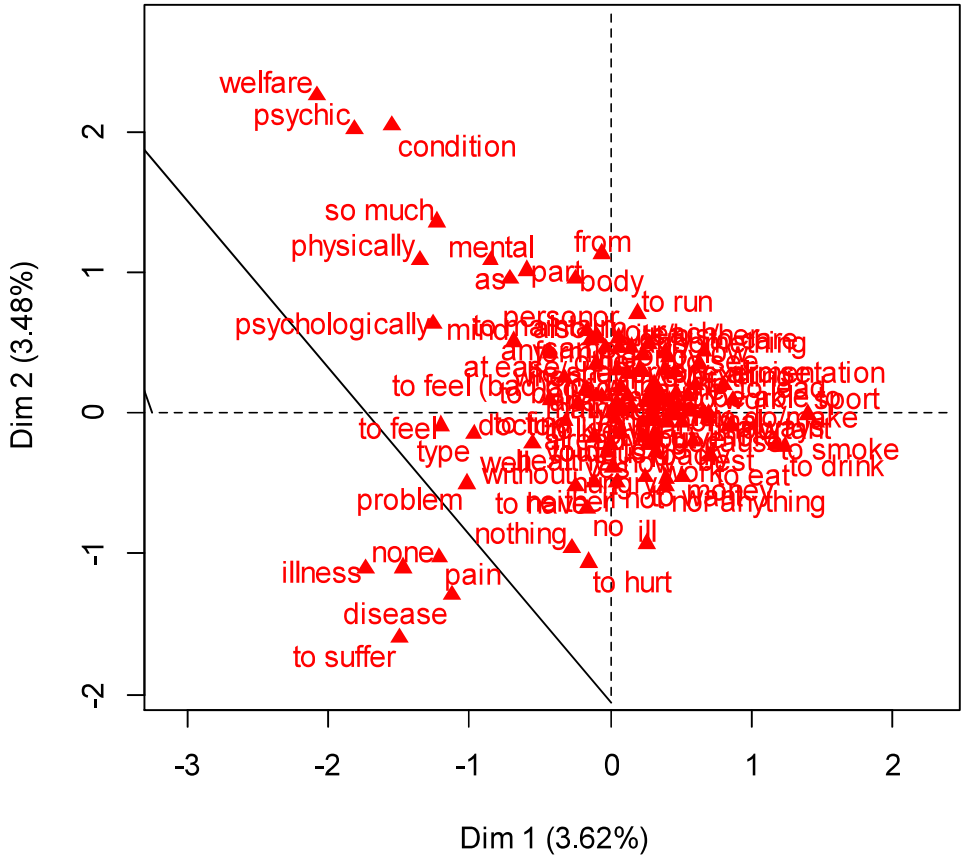
- The similarity between rows and columns is expressed in a totally symmetrical way (characteristic of CA which differentiates it from other principal components methods). Two rows are all the closer as they are frequently associated with the same columns and two columns are all the closer that they are frequently associated with the same rows.

Results CA: eigenvalues

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.2333100	3.619220	3.619220
dim 2	0.2243156	3.479693	7.098914
dim 3	0.1858920	2.883649	9.982563
dim 4	0.1753531	2.720163	12.702726
dim 5	0.1502916	2.331397	15.034123
dim 6	0.1433485	2.223692	17.257814
dim 7	0.1418055	2.199757	19.457572
dim 8	0.1348692	2.092157	21.549729
dim 9	0.1325759	2.056582	23.606311
dim 10	0.1293922	2.007196	25.613506
		.....	
dim 114	0.0070151	0.108822	100.000000

Results CA: first factorial plane





Results CA: highest contributions (1,2)

	Dim 1	Dim 2
physically	10.4460384	6.94756938
none	10.9328585	6.37469296
welfare	5.0371786	6.15360219
no	0.5879697	10.53125665
disease	4.5966688	6.36282802
psychic	4.1886973	5.36576630
condition	3.3205504	5.97001380
to feel	8.6498820	0.04843705
to have	0.3388698	6.13595298
nothing	0.3892418	4.96543895

From text to a word frequency table

PCA. This week, we have for you three videos which together present the main details of principal component analysis. Principal component analysis is a set of tools which allow us to study and visualize large data sets. We will present the method from a theoretical as well as a practical point of view….

CA. This week, we have for you 5 course videos on correspondence analysis. In the videos, we will see the following: first, we start by describing the data, giving a little notation, and considering questions to ask when running correspondence analysis. We'll see that the main point of correspondence analysis is studying the links between pairs of qualitative variables….

MCA. This week, we have four videos for you on multiple correspondence analysis, MCA for short. We'll have a look at the main features of the method, using a specific example to guide us along the way. The videos look at the following things. First, we describe the types of data MCA can be used for. With this data in mind, we will look at what our goals are, and what issues we may have….

Clustering. This week, we're going to look at clustering methods, including hierarchical clustering, and a partitioning method called k-means. The course videos for this week get into the following things: After a brief introduction on data of interest for clustering, and the goals of clustering, we are going to have a look at some general principles of clustering, a… particular, hierarchical …

1852 rows

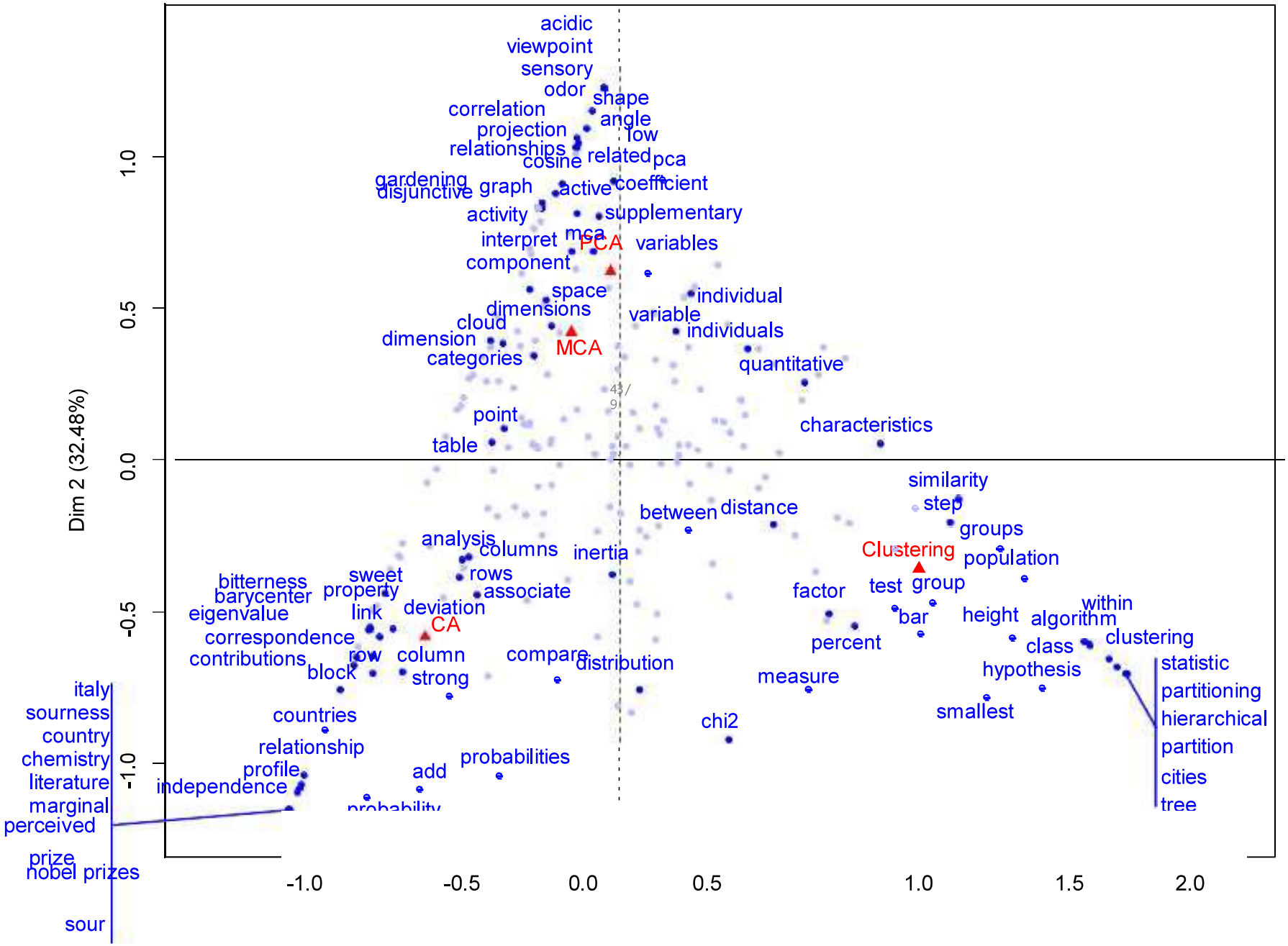
	PCA	CA	MCA	Clustering
able	2	1	1	2
above	0	0	0	1
absolute	1	0	0	5
absolutely	0	1	0	0
acceptable	0	0	0	1
access	1	0	0	0
accident	0	0	1	0
accord	0	0	1	0
according	3	0	2	0
account	0	0	2	1
…	…	…	…	…

Pre-treatment steps

- remove connecting words like : for example, then, therefore, and, etc.
- group words with the same root or the same conjugations together (e.g., reduced, reduction, reduces)
- group singular and plurals together
- remove words used nine times or less

$\Rightarrow$ 246 words, $n = 8821$ total occurrences

Simultaneous representation of methods and words



Interpreting the results

Terms exclusive to certain methods are superposed

1st axis (inertia = 0.35) :

clear division between the factor methods and clustering

specific words used in factor analysis are to the left

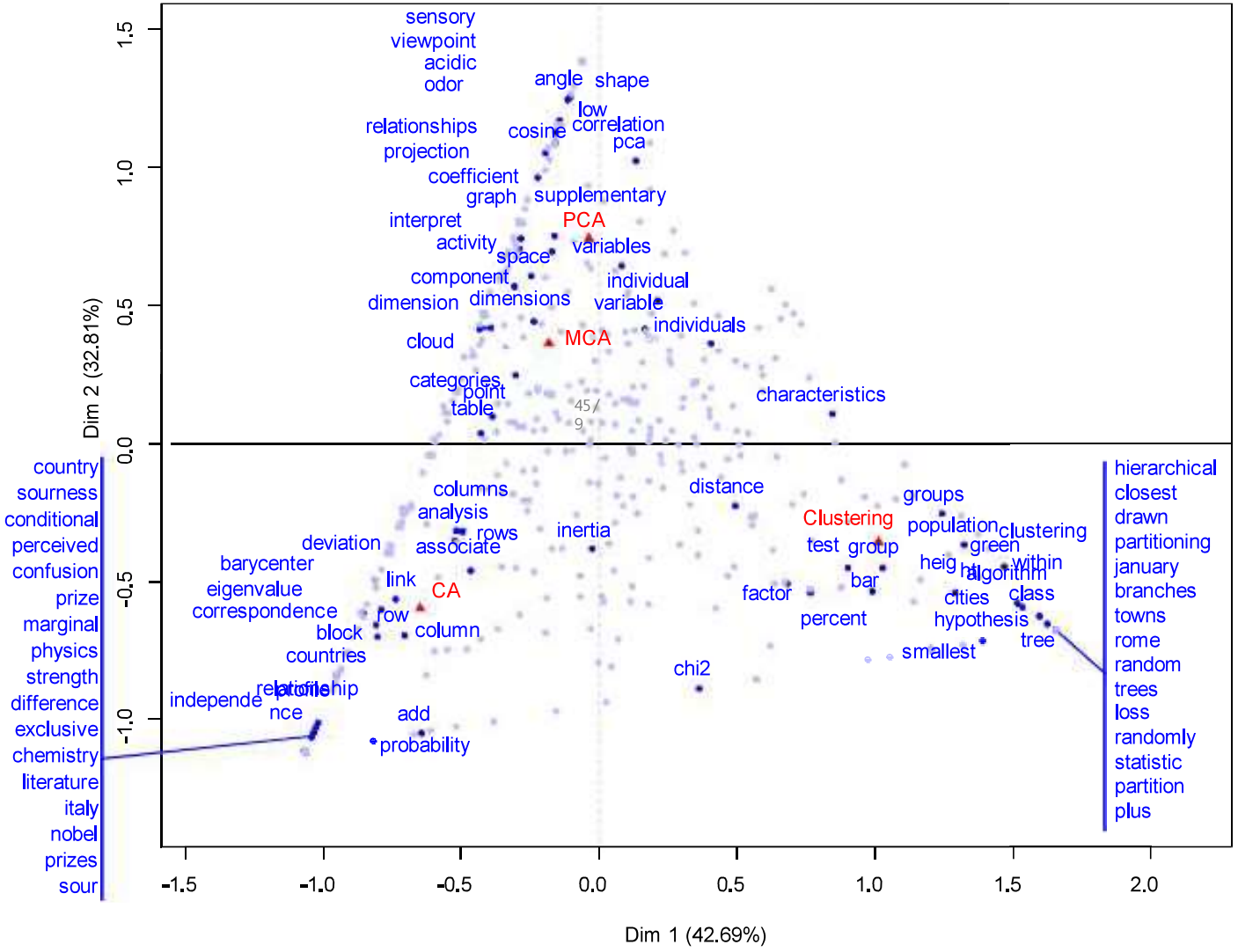
specific words used in clustering are to the right

2nd axis (inertia = 0.26) :

separates the 3 factor analysis methods

Stability of results with resect to cut-off

Words included : used at least 5 times



⟹ Stable representation

Bibliography

For more information about CA applied on textual data, have a look at this video:

https://www.youtube.com/watch?v=62N0JMf5hOs

Husson F., Lê S. & Pagès J. (2017) *Exploratory Multivariate Analysis by Example Using R* 2nd edition, 230 p., CRC/Press.

Benzecri, J. P. (1973). Analyse des Donnees. Dunod.

Benzecri, J. P. (1981). Pratique de l'analyse des donnees: Linguistique & lexicologie, volume 3. Dunod.

Escofier, B. and Pages, J. (1988). Analyses factorielles simples et multiples. Dunod.

Lebart, L., Salem, A., and Berry, L. (1998). Exploring Textual Data. Kluwer Academic Publishers.