A visual analysis of apartment renting in Paris

LF
UCBL Lyon1

SM
UCBL Lyon1

Abstract

Everyone is interested in renting an apartment in Paris for it is far from being an easy task. And that is what makes the subject worth studying. As visual analysis of this latter could bring some insight on the matter. That is what this work intends to do. It consist on an exploration of apartment renting in Paris through visual analysis. In other words the objective is to extract valuable information from the thousands of ads available on the net by visualising how they are written, and how they relate to eachother.

Index Terms: K.6.1 [Management of Computing and Information Systems]: Project and People Management—Life Cycle; K.7.m [The Computing Profession]: Miscellaneous—Ethics

1 Introduction [one page]

By exploring the data craweled from a well known French Ads Website, a first study was made on words and bigrams caracterising each renting ad. This led to a representation based on the vocabulary of the ads.

The purpose of this visualization is not just for data scientists interested in the subject who would like to extract some value from clustered adds.

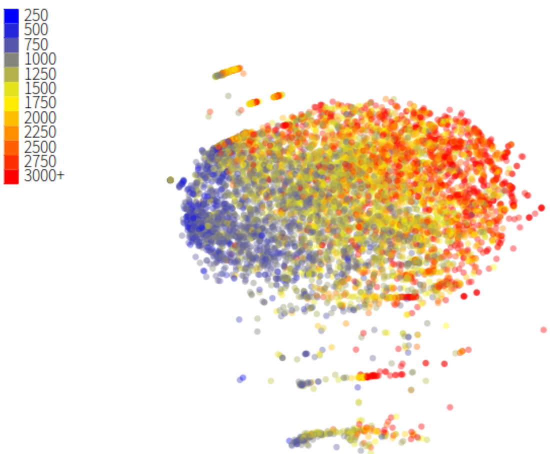


Figure 1: tsne 100

It is also helpful for anyone looking for an appartment. In fact, this could be used as a recommandation tool that empowers the user with an advanced recommanded tool. The idea is to let the user explore the ads as he or she usually does. The first interesting link they find gives them the chance to narrow their search to similar ones that are close in the representation space. This could be implemented as a browser's addon used in the Ads Website that the user hits feeding it with the ad's URL. The addon then sends a request with that URL as an argument. The page gets crawled, all the valuable text extracted, and transformed to the representation space. It then gets plotted and highlighted in the visualisation, giving the user the possibility to explore the space around it.

One can think that ads for different classes of appartments (small cheap ones, expensive big ones, very luxurious onces ...etc) differ in the vocabulary they use. A visualisation of caracteristics of words used in ads was made, the results of wich supported this assumption. It will be explained in further details in what will follow.

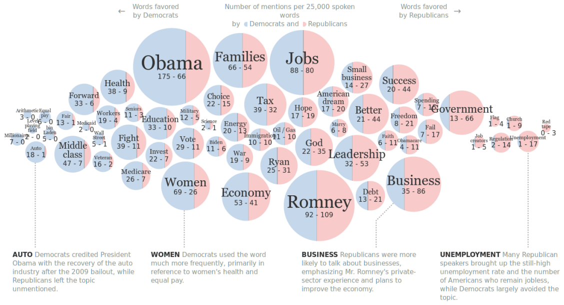
This paper is organised as follows. The next section will introduce some related works this project was based on. Then a project description will be given with all the steps that led to the final visualisation. The fourth part includes a discussions about the results and observations that were made. To finally end with a conclusion.

2 Related Works [one page]

2.1 Visualizing words used at the National Conventions 2012

The first visualisation of Words was inspired from the work of MIKE BOSTOCK, SHAN CARTER and MATTHEW ERICSON called **Visualizing words used at the National Conventions 2012** [1] that appeared in The New York Times in september 2012. The words were represented in bubbles.

The objective was to compare words used by democrats and republican in their speachs. A tool was also implemented for the users to add words of their own to see which party's speakers used it the most. It gives also the possibility to clic on a word and see the context in where the word appeared, which made it very interactive.



AUTO Democrats credited President Obama with the recovery of the auto industry after the 2009 bailout, while Republicans left the topic unmentioned.

WOMEN Democrats used the word much more frequently, primarily in reference to women's health and equal pay.

BUSINESS Republicans were more likely to talk about businesses, emphasizing Mr. Romney's private-sector experience and plans to improve the economy.

UNEMPLOYMENT Many Republican speakers brought up the still-high unemployment rate and the number of Americans who remain jobless, while Democrats largely avoided the topic.

Figure 2: At the National Conventions, the Words They Used [1]

2.2 NLP

In order to extract valuabale data, some Natural Language Processing tasks were implemented. In order to achieve that, python's Natural

Language Toolkit library was used [5]. It has predefined tokenizers, a list of stopwords in multiple languages including French, but also a premade tool for calculating frequency of words and bigrams.

2.3 Tsne

T-Sne (t-Distributed Stochastic Neighbor Embedding) is particularly well suited for the visualization of high-dimensional data. It is the technique that was used to generate the visulization above from the binary encoding of each ad.

3 Project Description [two pages]

3.1 Getting the data

To build the dataset, a creawler was coded to go throught the ads in the source website and store them in a JSON file. The file contained:

- The Title of the ad
- Its description
- A set of Tags like describing thinks like the floor in which the appartment was.
- The Price of the appartment
- Its address
- Its surface
- ...etc

After Extraction of the their descriptions as raw text, some preprocessing was done help with Natural Language Toolkit python library. The processing went as follows:

- Tokanization: words segmentation. (*Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens , perhaps at the same time throwing away certain characters, such as punctuation.* [4])
- Cleaning the text: getting rid of stopwords and numbers;
- Extracting words and tags frequency;
- Extracting bigrams frequency (sequences of two words);
- Computing average prices and areas for each word (average price and area of all ads it appears in);
- Storing everything in a pikle file.

The data was usued to build a descriptive visualisation.

3.2 A first visualisation

A first visualisation of the words used was made. This helps see notice some interesting patterns. For example:

- The ads of very expensive apartments tend to be written in english more often.
- Cheap ads were more explicit about basic housing needs.
- Expensive ads make use of the plural form of ”salle” (the french word for ”room”) a lot more often.
- ... etc

In the following Visualisation, the frequency of each word is represented by its radius and the color represents the average price of the ads it is associated with.

The valuable insights gotten from this visualisation led to the conclusion that the vocabulary used in an ad can be very different depending on the price tag. In other words, different ads used a different vocabulary to describe differenet type of appartments.

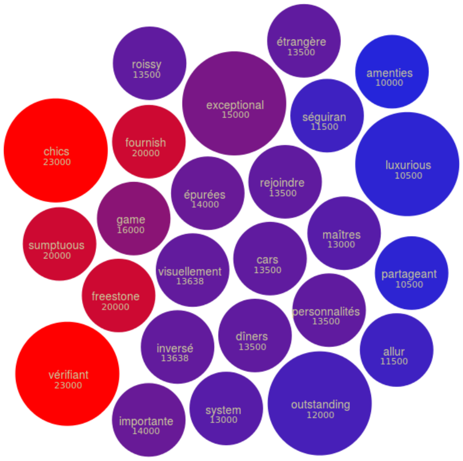


Figure 3: bubbles visualisation

3.3 Bag-of-words modeling

Encoding the ads with a bag-of-words model is the choice that was made for the next visualisation. This encoding was achieved through a further processing of the extracted data.

- The words, bigrams and tags were regroupped in one list.
- The list was sorted with respect to the frequency of the appearance of its elements in the ads.
- A sublist of the elements that appeared more than a hundred (100) time was chopped.
- The later list was rechecked manually by the authors. A vector of 800 words was retained.
- Then for each ad crawled contained in the dataset, a representation was made based on that vector. The ad has a value of 1 in the position of the words that appear in it, and 0 in that of the words that don’t.

3.4 Dimentionality reduction and plotting

For Dimentionality reduction, tsne (t-Distributed Stochastic Neighbor Embedding) was used. It is particularly well suited for the visualization of high-dimensional data. It is the technique that was used to generate the visulization above from the binary encoding of each ad.

At first a attempt was made to use a JavaScript implementation of tsne that is mentioned in Laurens van der Maaten’s website. But this latter was not based on an old version of the algorithm [2] the complexity of which was quadratic to the number of points. So the most suited was to use the scikit-learn implementation that was based on a more recent and much faster version of the algorithm [3].

- use an effective dimensionality reduction technique (t-Sne)
- add a little bit of colors (depending on the price)
- Here is another example of t-Sne (on the same dataset but with a different learning rate).
- alternatives: $[\text{viz}_m ds][viz_p ca]$

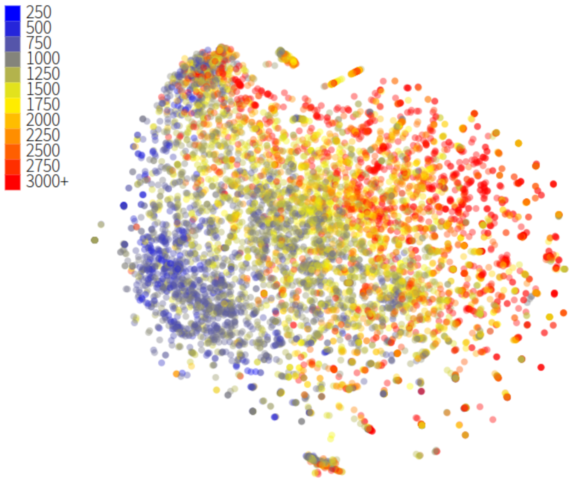


Figure 4: tsne 1000

4 Discussion [1/2 page]

- Can you see the color gradient? This shows how different the vocabulary used in the ads descriptions can be different depending on the value of the appartement.
- Can you see the small clusters? This shows how renting agencies tend to use the same template of ad description over and over again.
- Can you see that there are more clusters in the red areas than in the blue areas? This shows that expensive appartments tend to be managed by renting agencies more than cheap(ish) appartments.
- distance metrics

5 Conclusion [1/2 page]

In conclusion, ...

6 References

* [1] MIKE BOSTOCK, SHAN CARTER and MATTHEW ERICSON; Visualizing words used at the National Conventions 2012; The New York Times; september, 6th 2012. Link: http://www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html

* [2] van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data; Using t-SNE. Journal of Machine Learning Research 9:2579-2605, 2008.

* [3] L.J.P. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 15(Oct):3221-3245, 2014.

* [4] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schtze, Introduction to Information Retrieval, Cambridge University Press. 2008.

* [5] Bird, Steven; Klein, Ewan; Loper, Edward; Baldridge, Jason (2008). ”Multidisciplinary instruction with the Natural Language Toolkit” (PDF). Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics, ACL.

* MDS

* PCA

* About color choice?

Acknowledgments

Thanks to our tutors who gave us both freedom and guidance during this project.