Consumers visual associations with brandnames

Feshchenko, Ilya Vlasova, Anna Riabukhina, Daria
Dzunja, Dejan Sadri, Mohammad Ali

21/03/19

1 Introduction: visual associations

Our project is addresses mainly practical issues, it is about text processing techniques, in particular, topic elicitation methods, applied to a dataset based on a visual data. Visual associations elicitation is an extremely important task for business and marketing.

Another goal of our project is a deeper understanding of what topic is, and more importantly, what topic is for visual-based data. This question is only beginning to get unexplored. In Natural Language Processing there are two most common approaches for solving such task as topic modeling.

The first one is to use parametric models of the distribution of words across documents of a collection in order to capture the underlying structure of such a distribution. The existence of such a structure relies on certain assumptions. Our team's goal is to understand whether they hold for visual-based data.

The second approach is based on continuous space embeddings representation of words. Continuous space is constructed in such a way so that certain metrics as cosine similarity represented for pairs of different words something that a human being would interpret as closeness of words in meaning. In this setup, topic modeling may be solved as a clusterization problem in a continuous space. The success of this approach is not guarantied, neither is the success of the first one, because it relies on certain assumptions, too.

These two approaches typically work independently, and hence, may yield different results of topics modelling. Though the complimentarity of such ap-

proaches is observed for Natural Language Processing, it is not at all obvious that the same would hold for visual-based data.

In our work we test whether there is any contradiction, independence of complimentarity between these two approaches and underlying assumptions for visual-based data, so that we could better understand what can be considered as a visual association, or topic for visual-based data.

1.1 Visual associations: practical motivation

Let us briefly outline two of the main problems a proper visual associations elicitation algorithm may solve.

1.1.1 Economy boost: market inefficiency treatment

Firstly, businesses often do not know how to sell. If they do not know which visual associations with their product potential consumers may have, this typically results in no advertisements or irrelevant ads, for example, something erotics-related. Though such ads may be catchy and may prevent the entire collaps of a company, they tell nothing about the product itself. hence, since consumers remain uninformed, they are more likely to do poor product choices.

As a result, business' profit maximization is hindered, as well as consumer's utility maximization under poor product choice. There is an inefficiency in the market caused by incomplete information, partially caused by businesses' ignorance regarding potential consumer's visual associations.

1.1.2 Enterpreneurial interests: brand expansion failure prevention

Secondly, when enterpreneurs dream of new goals and try scaling their brand into a new market, sometimes they make strange decisions regarding the market/product choice for brand expansion, because the new product is too weakly associated with an old one familiar to consumers. As a result, they suffer considerable losses.

Such weak associations include visual ones. Thus, this problem can be done less severe if businesses check consumers' visual associations with their brandname before taking any expansion steps.

1.2 Visual associations: how?

As we have mentioned, we use text analysis techniques. Let is briefly outline the origin of the dataset and the means by which the visual problem is converted into a textual one.

There is a survey on the web asking people which brandnames they are familiar with, and then aking to drag into a collage pictures they may find corresponding to their visual assiciations with the familiar brandnames. It's in a way like a product review but instead of words people leave puctures. As the collages are collected, the pictures are went through a CNN API on the web assigning a set of corresponding tags to each of the pictures in a collage.

After the problem is converted from a visual into a textual one, different techniques for topic extraction may be applied, like Gaussian Mixture clusterization, Latent Dirichlet Allocation of Nonnegative Matrix Factorization, modeling different topics and yielding different sets of visual assiciations.

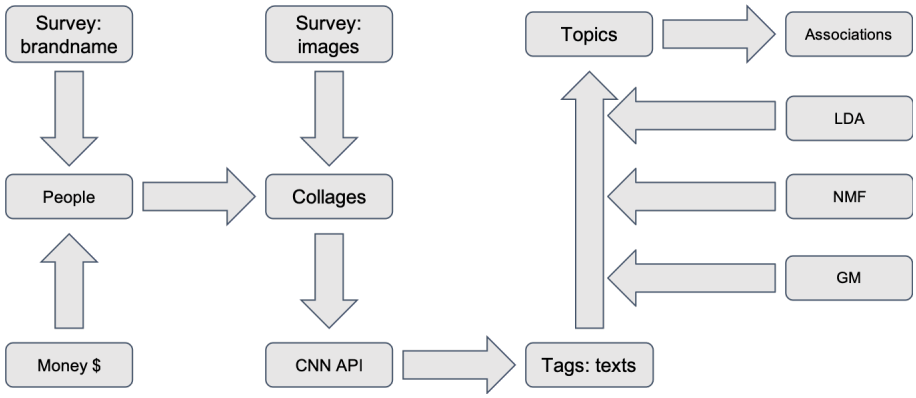

Figure 1: The origin of the Collages dataset and related workflow.

1.3 Visual associations: theoretical motivation

There is no surprise that topic modeling quality assessment technuques based on continuous-space word embeddings and based on words distribution among documents in a collection are in a way complimentary. For example, word2vec [7] is likely to assign the higher cosine similarities to a pair of words the higher they occur in a similar context. These chances obviously increase if a

pair of words often occurs in similar documents, so for textual data for topics understanding and modeling the existence of a complimentarity between embedding clusterization and such statistical distribution modeling as NMF or LDA is quite trivial. This argument may help to interpret and justify the existence of a link even though embeddings such as word2vec may rely mainly on the distance separating a pair of words, whereas LDA or NMF work under bag-of-words assumption.

The bag-of-words assumption of irrelevance of the words' order in a document indeed what we are dealing with in such visual-based dataset as Collages. Thus, embeddings are better not to be trained on Collages, and it brings up an intriguing question of whether a good pre-trained embedding may be used for destribution-based topics model quality assessment.

In contrast to datasets with a purely textual origins, for the visual-based data, like Collages dataset, a presence of a complimentarity between two topic modeling approaches represents a particular research interest, and would be an important discovery meaning that the optimization objectives of different in underlying assumptions methods may and should be optimizad simultaneously.

2 Related work

As we have mentioned, different embeddings clusterization techniques may be used, like K-means[5], which is nothing but a Gaussian Mixture [1] with a variance set to zero, so GM is better in a sence that it is a more general setup which in addition addresses hard assignment flaw K-means has (In general, there is nothing wrong about same words occuring in different topics). KNN[4] may also work, and' what's important, it can work with cosine similarity metric.

As generative text models working with parametric distributions, capturing some underlying distribution structure, we can take LDA [2], or NMF [8]. Interestingly, we can use continuous-space embedding-based metrics for such models quality assessment. Such an idea is very useful for our goal to understand whether there is a complimentarity between two common topic modeling approaches for visual-based data.

In case a complimentarity is proven, a relatively new method of GMM-LDA [3] model is reasonable to apply to fit this complimentarity best and to take into account both embeddings and distribution strustures.

collage_id	description	surveyStart_time	imageID	imageLayerNumber	searchWords
14	some refreshment pictures here to make feel ha...	4/4/2017 22:54	21911, 31554, 43877, 83298, 32166, 1565, 55152...	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	fresh
47	It's relaxing, refreshing and cool.	4/4/2017 22:58	82443, 13083, 33779, 42473, 2724, 73500, 31150...	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	refresh, enjoy, river, thirst, neck, sun, glac...

Figure 2: Collages dataset

	asin	helpful	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
0	616719923X	[0, 0]	4	Just another flavor of Kit Kat but the taste i...	06 1, 2013	A1VEELTKS8NLZB	Amazon Customer	Good Taste	1370044800
1	616719923X	[0, 1]	3	I bought this on impulse and it comes from Jap...	05 19, 2014	A14R9XMZVJ6INB	amf0001	3.5 stars, sadly not as wonderful as I had hoped	1400457600

Figure 3: Amazon dataset

3 Dataset Description

3.1 Collages

Collages is a dataset that contains lots of data from the survey ($\approx$ 150 different fields), like images people use, search words they type to find images and summaries they white to their collages (fig. 2). There is more of textual data, yet our main goal is to explore visual associations, so we only use tags generated by CNN API for the purity of the experimental setup.

Such a visual-based dataset is quite rare, and that is one of the reasons it represents such a high research interest. That is exactly why we perform the most of the methods at this dataset, and we had use for datasets' variety a dataset of a usual textual origin, which is a dataset of Amazon product reviews.

3.2 Amazon product reviews

This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.

This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).(fig. 3)[6]

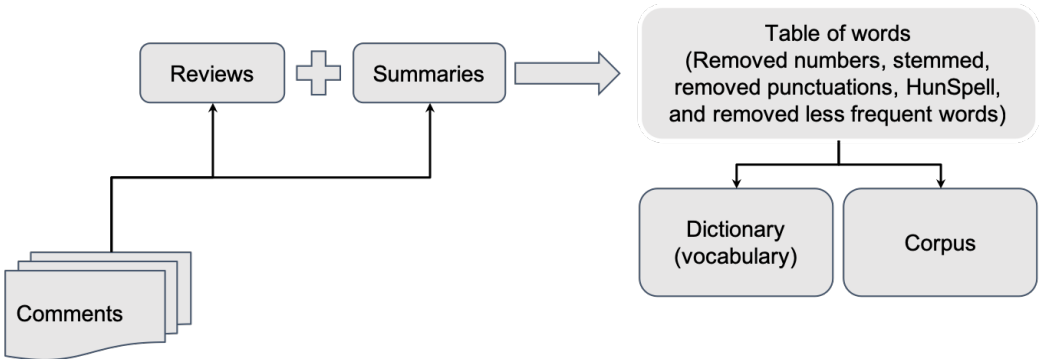

Figure 4: Preprocessing procedure

4 ML Methods

4.1 Preprocessing

4.1.1 Amazon

For Preprocessing of the Amazon reviews dataset we used all comments and summary entries. The provided comments data passed into the function to get processed and extracted words of each comments. The function consists of: Removing punctuations, Tokenize strings, removinig numbers, removing stopwords, stemming words and, removing words with wrong dictation.(fig. 4)

4.2 Models

4.3 Why do we use for topic modeling collages as documents and not brandnames

Firstly, we use collages as documents because the assiciations may depend on user. Such a model may be a step towards a more ambitious model making predictions, for instance, of review scores based on collage associations, of making profit maximization recommendations depending on what different users can afford. The idea for such ambitious tasks is to distinguish associations among different users and not just among brandnames.

What is also important is that the models should remain applicable when a new brand is incorporated into the review, so, leaving brandnames out of

consideration here is a means to prevent overfittiing so that the model could better meet new brands' data.

4.3.1 Collages

To generate the Collages dataset the survey is used. The best way to understand the dataset is to participate in its generation, to pass the survey thet gathers Collages data

To convert the visual problem into a textual one the clarifai CNN API is used.

In context of our work, especially its research aspect, it is crucial to use embeddings. As a part of pre-procerring, we use TensorFlowHub's Universal Sentence Encoder

4.4 Metrics

We use NMF, LDA that are capturing the distribution of tags among collages. The bigger the number of topics the better these models fit to data in a distributional sence. In order to test the presence of a complimentarity between distribution fitting and embedding-space fitting approaches it is reasonable to let these methods go as usual and iterate over the number of visual associations, but to plot a metric that represents continuous-space embedding-based fit quality.

4.4.1 Cosine Similarity Ratio

We introduce such a metric that shows how homogeneous are a certain number of different words within one topic on average, and by how much times this homogeneity in a sence of cosine similarity exceeds the average cosine similarity among different topics.

For a given topic elicitation model with topics T, the following method considers each of the topics $t \in T$ as a set of $n\_top\_words = n_{tw}$ words such that $\forall w \in t$ the probability to occur in a topic t is $\mathbf{P}(w|t)$, and the method computes a CDR metric (Cosine Distance Ratio) for the model with topics T defined as follows:

$$CDR(T) = Cosine\ Distance\ Ratio(T) = \frac{\frac{1}{\|T\|}\sum_{t\in T} ITS(t)}{\frac{1}{T(T-1)}\sum_{t_1\in T, t_2\in T, t_1\neq t_2} ETS(t_1, t_2)},$$

where

$$ITS(t) = Internal\ Topic\ Similarity(t) =$$
$$= \frac{\sum_{w_1,w_2 \in t, w_1 \neq w_2} \mathbf{P}(w_1|t)\mathbf{P}(w_2|t)\langle w_1, w_2\rangle}{\sum_{w_1,w_2 \in t, w_1 \neq w_2} \mathbf{P}(w_1|t)\mathbf{P}(w_2|t)}$$

and

$$ETS(t_1, t_2) = External\ Topic\ Similarity(t_1, t_2) =$$
$$= \frac{\sum_{w_1 \in t_1, w_2 \in t_2, w_1 \neq w_2} \mathbf{P}(w_1|t_1)\mathbf{P}(w_2|t_2)\langle w_1, w_2\rangle}{\sum_{w_1 \in t_1, w_2 \in t_2, w_1 \neq w_2} \mathbf{P}(w_1|t_1)\mathbf{P}(w_2|t_2)}$$

are expected cosine similarities of different words, within a singular topic t, and across two different topics t_1 and t_2, respectively, and $\langle \cdot, \cdot \rangle$ is cosine similarity; for the embedding algorithm we are using $\|w\| = 1\ \forall w$ known to embedding, and hence $\langle \cdot, \cdot \rangle : \mathbf{R}^{512} \times \mathbf{R}^{512} \to [0; 1]$.

The higher the metric, the more homogemeous in a cosine similarity sence are different words on average within a singular group, in comparison with similarity across different groups.

In theory it may take values in $CSR(T, n_{tw}) \in [0, +\infty)$, whereas the expected values and the values in in practice are bigger than 1 (which corresponds to a random topic elicitation): $CSR(T, n_{tw}) > 1$.

4.4.2 CSR plots

We plot the line graphs of CSR in *n_topics* for NMF (fig. 5) to test whether it grows as distribution fits better. Our methods in the code allow to use the saved models to plot these graphs for a various numbers of most frequent topic words $CSR(T, n\_top\_words)$, so it's impossible to fit them all into report.

For confidence intervals estimates we use corrected sample variance for CSR computed for samples of model trained on datasets obtained by bootstrap resampling.

The line graphs of CSR for all *n_top_words* parameters exhubit sustainable and continuous growth.

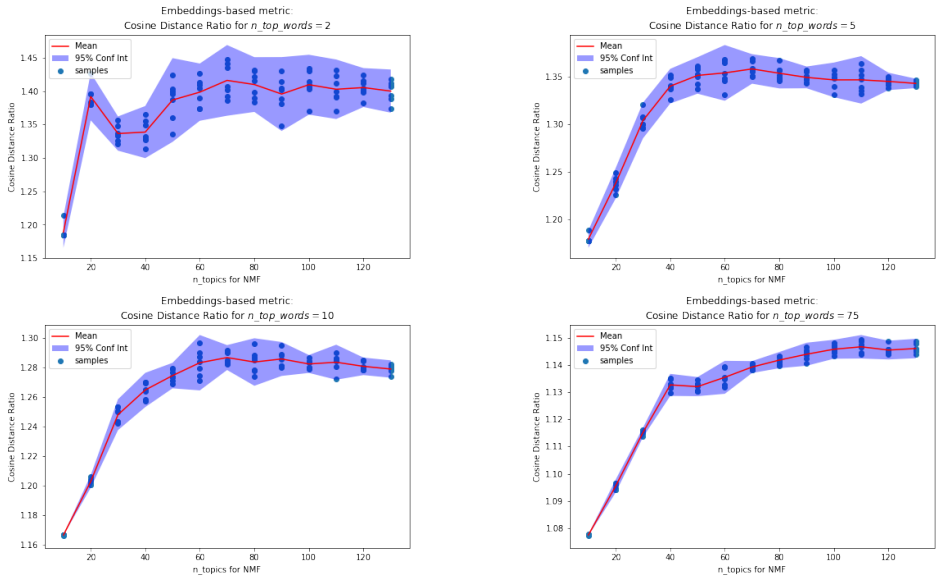

Figure 5: CSR

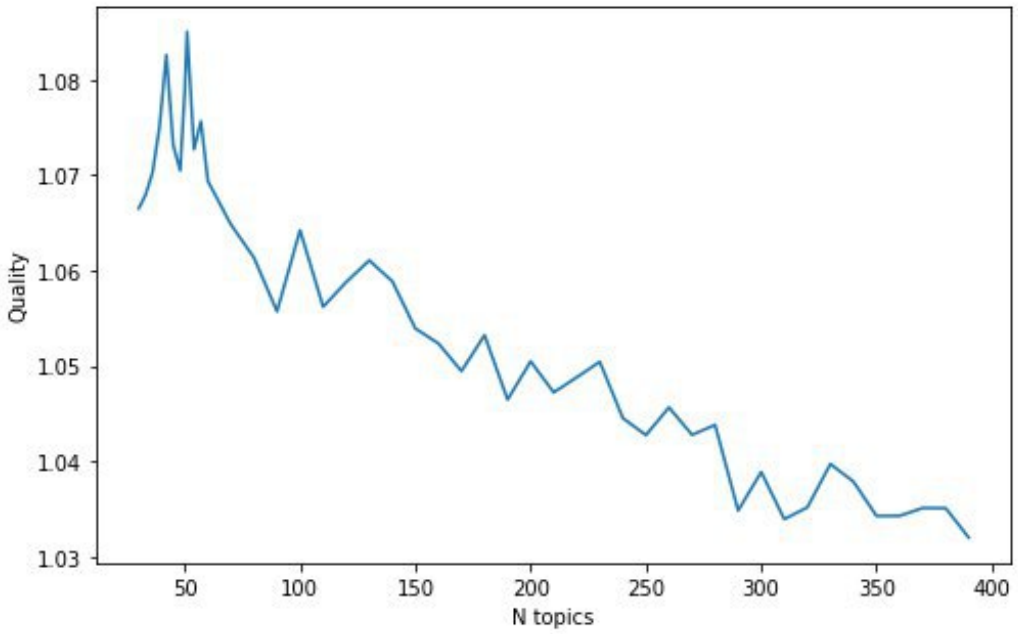

Figure 6: CSR for LDA

4.5 Visualization: PCA

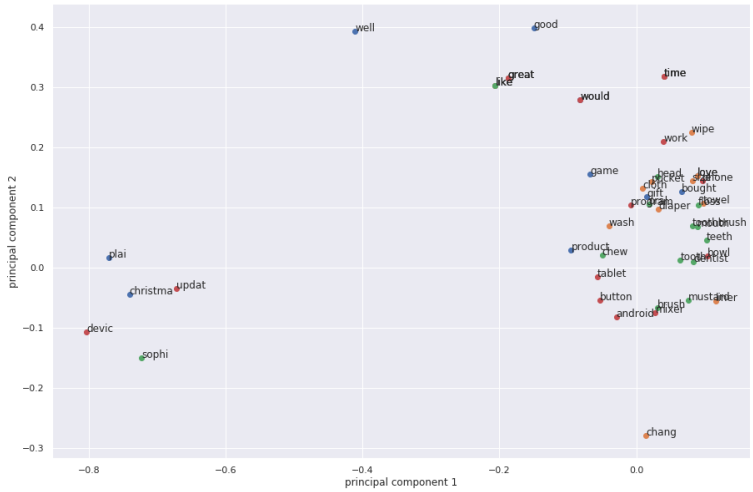

Figure 7: PCA

4.6 Visualization: word clouds

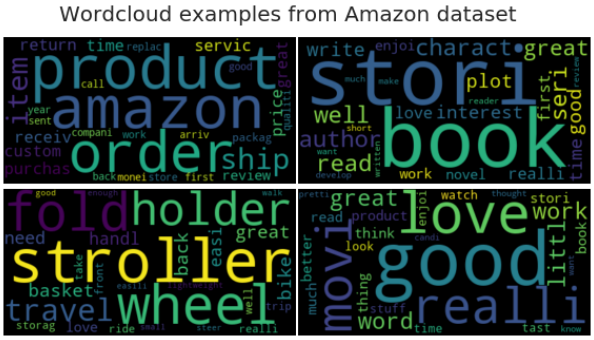

Figure 8: Word Clouds

4.7 Perplexity of LDA model

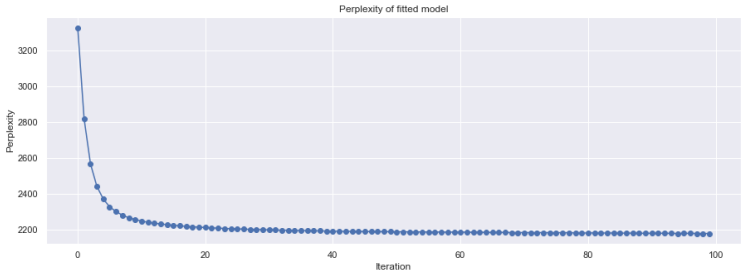

Figure 9: Perplexity of LDA model for 100 iteration

5 Results and conclusion

Apparently from the line graphs, as *n_topics* goes up and the quality of the collection-distribution capturing goes up, the metric based on TensorFlow Hub Embedding gradually increases as well, which indicates the complimentarity of embeddings space clustering and generative parametric text models. However, there is a saturation level or even global optimas, which may in a sence indicate a limited nubmer of topics, and also likely to be caused by the fact that cosine similarity and distribution fit likelihood are not the same thing and their connection is limited.

For LDA for Collages dataset there is a maxima of CSR, which means there is a room for further research of the complimentarity for visual data, or that there is an optimal number of topics.

Both LDA and NMF for Collages stop improving significantly after $n\_topics = 50$

The innovative result is that our work highlights high perspectives of learning methods working with both collections of documents and embeddings space, such as GMM-LDA, in visual-based text analysis application.

We have succeeded to obtain interpretable topic for both Amazon dataset and Collages dataset.

References

[1] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov

models. *International Computer Science Institute*, 4(510):126, 1998.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 795–804, 2015.

[4] James M. Keller ; Michael R. Gray ; James A. Givens. A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15, 1985.

[5] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[6] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *CoRR*, abs/1602.01585, 2016.

[7] Xin Rong. word2vec parameter learning explained. *CoRR*, abs/1411.2738, 2014.

[8] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

6 Contributions

6.1 Feshchenko, Ilya

Team leader, project ideas: topic modeling methods, visualizations, CSR, spellcheck and stemming.

Coding: early LDA trials; PCA visualization, CSR; Collages corpus; bootstrap; CSR for NMF; topic naming script.

6.2 Vlasova, Anna

Project theoretical guidance through generative probabilistic text modeling. Found an article and matlab GitHub repository on GMM-LDA.

Coding: K-means, CSR for LDA.

6.3 Riabukhina, Daria

Coding: preprocessing of Amazon and Collages; Embeddings; project report

6.4 Dzunja, Dejan

GitHub, AWS EC2, code structuring, project meetings planner

Coding: LDA; WordClouds visualization; Stemming; Spellcheck

6.5 Sadri, Mohammad Ali

Public relations (mailed to Amazon to get data).

Preprocessing: Amazon dataset (Tokenize, Stopwords, ...).

Coding: Download tool for dataset, Constructing Amazon corpus, LDA, Compare LDA performance from different packages.