Text Analytics on Start-up Descriptions

Olivier Parisot, Patrik Hitzelberger, Yoanne Didry
Luxembourg Institute of Science and Technology
41 rue du Brill, Belvaux, Luxembourg
email: olivier.parisot@list.lu

Gero Vierke, Helmut Rieder
InfinAIt Solutions SA
2a Ennert dem Bierg, Sandweiler, Luxembourg
email: gvierke@infinait.eu

***Abstract*—In order to analyze the descriptions of various active start-ups, we have developed a web application to retrieve and to analyze available textual data about them. The tool aims at extracting the frequent topics and applying semantic similarity analysis to the start-up descriptions.**

I. Introduction

Nowadays, the creation of start-ups is a major political objective in many countries, trying to generate technological innovation and fast-growing businesses. This phenomenon is widely studied in the literature from an economical perspective, looking e.g. at issues regarding intellectual property [1].

Given the importance of start-up companies, public and private players are however interested in actual information about individual companies and the overall market [2].

To this end, we propose to apply advanced text analytics technologies on online available data by taking advantage of the richness of the API ecosystem [3].

In this paper, we present a software to retrieve and analyze start-up descriptions. We have selected and integrated various components to apply state-of-the-art text mining techniques.

II. Prototype

We have realized a web application with the Grails framework [4]: the *business-logic* layer is implemented in Java & Groovy, and HTML5/Javascript is used for the user interface.

The prototype is based on micro-services architecture principles [5]: this kind of approach allows to cleverly integrate heterogeneous components by managing availability issues and queries caching. Then, various third-parties components were integrated into the prototype (Table I): *a*) Open source libraries were directly integrated into the application. *b*) Remote APIs are invoked during the prototype execution.

TABLE I
Third-parties components.

Name	Type	Purpose
Y Combinator	Remote API	Data retrieval
Apache Tika	JAVA library	Language detection
Apache OpenNLP	JAVA library	Language processing
AlchemyAPI	Remote API	Concepts extraction
DISCO	JAVA library	Similarity analysis
MALLET	JAVA library	Topic modeling
Highchart	Remote API	Data visualization

Firstly, we have developed a data loader to retrieve the descriptions of the start-ups by getting the publicly available data of '*Y Combinator*', a well-known and successful start-up fund started in 2005 [6]. For each listed start-up, we extract the name, the short description and the URL.

Secondly, these textual data are analyzed as follows:

- The language of the description is detected with Apache Tika, by using a classification algorithm [7].
- The stop words are then ignored by using the Snowball tool's word lists, and a stemming process is realized [8].
- For all the start-up descriptions, the main concepts are extracted via AlchemyAPI [9], a remote service for text mining that is mainly based on *deep learning* [10].
- Moreover, the main topics are extracted via MALLET, a topic modeling library [11].
- A semantic similarity analysis is realized on descriptions with DISCO, a multilingual corpus-based library [12].

Finally, the results are presented with the following Javascript libraries: jQuery [13] and HighChart [14].

III. Main topics discovery

The prototype provides a first module to extract and visualize the terms and topics that are present in the start-up descriptions. This component aims at discovering the main tendencies in start-up activities. As an example, we suppose that the descriptions are widely filled by technical terms like *application* or *mobile*. However, we think that data analysis will help to discover the less obvious but important topics.

Therefore, our prototype provides a module to generate tag clouds to simply show frequent terms [15]. Then, the MALLET engine is applied to regroup the terms into clusters, depending on their co-occurrence. Finally, the AlchemyAPI service helps to extract the important concepts from the text by applying *deep learning* [9], [10] (these concepts can be composed of several words like *social media*). Mixing tag clouds, clusters and main concepts helps to refine the results.

IV. Semantic similarity analysis

The second feature of the prototype aims at applying semantic analysis to allow the user to execute queries on a list of start-up descriptions (this approach is known as '*Concept based query expansion*' in the literature [16]). Instead of searching the texts that exactly match, we propose to return the descriptions that are *semantically close* to the searched terms.

Thus, we provide an interface where the user can enter queries as word or a sentence. Based on this input, the

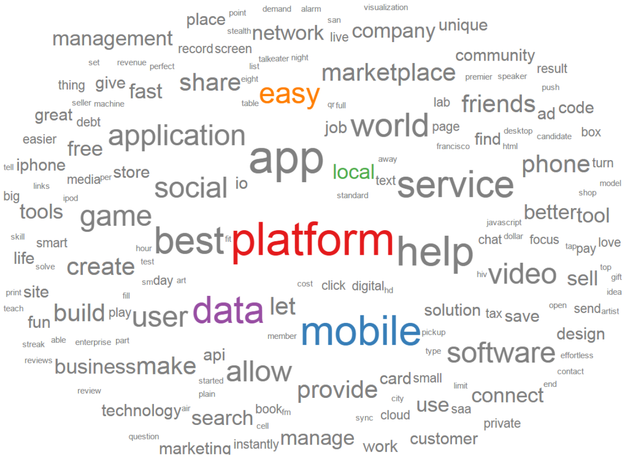
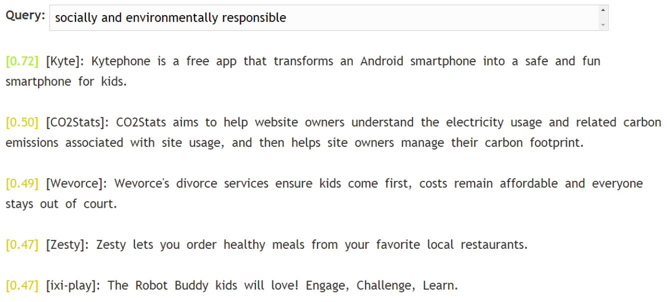


Fig. 1. Tag cloud showing the words that are frequently present in start-up descriptions. The colored words are the main topics found with MALLET.



Query: socially and environmentally responsible

[0.72] [Kyte]: Kytephone is a free app that transforms an Android smartphone into a safe and fun smartphone for kids.

[0.50] [CO2Stats]: CO2Stats aims to help website owners understand the electricity usage and related carbon emissions associated with site usage, and then helps site owners manage their carbon footprint.

[0.49] [Wevorce]: Wevorce's divorce services ensure kids come first, costs remain affordable and everyone stays out of court.

[0.47] [Zesty]: Zesty lets you order healthy meals from your favorite local restaurants.

[0.47] [ixi-play]: The Robot Buddy kids will love! Engage, Challenge, Learn.

Fig. 2. Start-up descriptions retrieved for the query: '*socially and environmentally responsible*': a semantic similarity score is given for each result (between 0 and 1 – 0 for low similarity, 1 for high similarity).

prototype applies the semantic analysis engine to list the companies for which the descriptions are semantically similar.

Additionally, the prototype generates a Multidimensional scaling (MDS) projection to represent the similarity of the textual data in a 2D plot (as suggested in [17]). In practice, the distance between each text is obtained by adapting the DISCO similarity function. As a result, the user can asses the similarity of the projected texts with a user-defined term: it is visually highlighted with the *Colorbrewer* palettes [18].

V. Experiments

We have conducted data analysis on the descriptions of the 591 active start-ups currently listed by *Y Combinator* [6].

As a first illustrative example, we can see with tag clouds that the most frequent terms are a mix of technical words (*platform*, *data*, *mobile*), subjective adjectives (*easy*, *best*) and verbs (*help*, *provide*, *allow*) (Fig. 1). Even if the results seem obvious (listed companies are mainly related to IT), it helps finding technologies that are currently considered as innovative. This is confirmed by the AlchemyAPI analysis: most important concepts are *mobile apps* or *data science*.

To find *less obvious* topics, we tried to find out the important non-technical concepts. We used our search module, and we found that various companies are targeting IT-specific markets like *social networks* or classical business domains like *finance* (Fig. 2). We built various MDS projections and we found that a lot of start-ups are related to the '*health*' domain (Fig. 3).

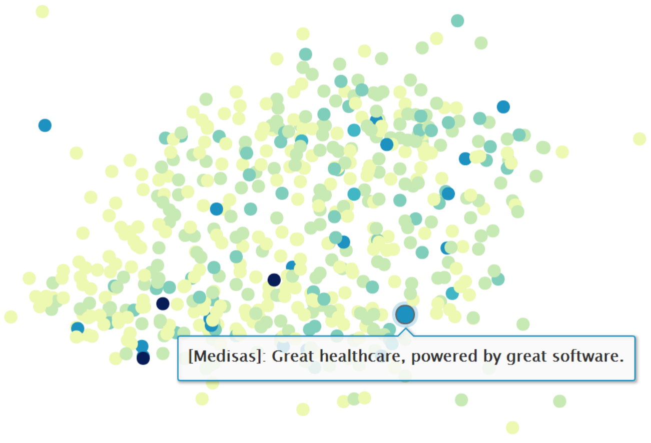


Fig. 3. MDS projection showing the semantic distance between the start-up descriptions. The color represents the relatedness with the '*health*' term [18].

VI. Conclusion

We have developed a prototype combining text analytics and data visualization to build an overview of start-ups activities based on online available textual data about them. The platform aims at helping investors, entrepreneurs or jobs seekers to understand the real activities of these companies. In future, we plan to analyze projects presented on *crowd-funding* platforms.

Acknowledgment

The project is supported by a grant from the Ministry of Economy and Foreign Trade of Luxembourg (RDI law).

References

[1] D. Di Gregorio and S. Shane, "Why do some universities generate more start-ups than others?" *Res. policy*, vol. 32, no. 2, pp. 209–227, 2003.

[2] T. R. Stone, "Computational analytics for venture finance," Ph.D. dissertation, UCL (University College London), 2014.

[3] S. Zilis, "The current state of machine intelligence 2.0," https://www.oreilly.com/ideas/the-current-state-of-machine-intelligence-2-0, 2015.

[4] P. Ledbrook and G. Smith, *Grails in Action*. Manning Publ. Co., 2014.

[5] D. Namiot and M. Sneps-Sneppe, "On micro-services architecture," *Int. Journal of Open Information Tech.*, vol. 2, no. 9, pp. 24–27, 2014.

[6] "Y Combinator company list," http://yclist.com/, 2016.

[7] C. Mattmann and J. Zitting, *Tika in action*. Manning Publ. Co., 2011.

[8] M. F. Porter, "Snowball: A language for stemming algorithms," 2001.

[9] J. Turian, "Using alchemyapi for enterprise-grade text analysis," Technical report, AlchemyAPI (August 2013), Tech. Rep., 2013.

[10] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, 2006.

[11] A. K. McCallum, "MALLET," 2002, http://mallet.cs.umass.edu.

[12] P. Kolb, "DISCO: A multilingual database of distributionally similar words," *Proceedings of KONVENS-2008, Berlin*, 2008.

[13] E. Sarrion, *jQuery UI*. " O'Reilly Media, Inc.", 2012.

[14] J. Kuan, *Learning Highcharts 4*. Packt Publishing Ltd, 2015.

[15] J. Sinclair and M. Cardew-Hall, "The folksonomy tag cloud: when is it useful?" *Journal of Information Science*, vol. 34, no. 1, 2008.

[16] Y. Qiu and H.-P. Frei, "Concept based query expansion," in *SIGIR conference*. ACM, 1993, pp. 160–169.

[17] Á. Blanco and M. Martín-Merino, "A partially supervised metric mds algorithm for textual data visualization," in *IDA*, 2007, pp. 252–262.

[18] M. Harrower and C. A. Brewer, "Colorbrewer. org: an online tool for selecting colour schemes for maps," *The Cartographic Journal*, 2003.