SMART

Social Media Analysis and Response Tool - Data Science Consulting

# Basic goals

1/ Define the most relevant data from social media (Twitter, LinkedIn, Facebook, Blogs, Youtube commentary, Google+, Amazon, Yelp); news (AFP NewsML, Guardian, Financial times; Wall Street Journal, Times, TV News report), scientific publications (Jstor, Google Scholar, Web of Knowledge, specific journals); Wikipedia; TED conferences; GoogleTech; Internet videos; laws, official statement from Congress, White House & Courts (C-span), data from a company[1](#_ENREF_1).

2/ Download and classify the data in real time using a NoSQL (non-relational, schema less, cluster friendly) approach.

3/ Use an algorithm to analyze the data. By default: topic modelling based on partially supervised learning model (“Latent Dirichlet Allocation”, Bley, Ng and Jordan, J. of Machine Learning Research **3** (2003) 993-1022). Dynamic (evolution in time), correlated (interaction between topics), supervised (prediction) models. Clustering for short messages with only one topic.

4/ Define the information needed (e.g. ‘number of ice cream sold’ + ‘economic situation’ + ‘store brand FB Likes’); choose the topics, extract and analyze the information (three quantities increase and are correlated) that will permit to act or respond in order to maximize (or minimize) the desired parameters (be ready to sell more ice cream).

5/ Quantify the effect of the response (did customers like the new flavor? Is the waiting time now too long?).

6/ Predict trends by defining relevant data (if other *gelateria* expand after costumers liked the new ‘cloud flavor’, maybe you should sell it too).

# Challenges

1/ Data acquisition (Spark?), storage.

2/ NoSQL

3/ Topic modelling.

**DO**: Allow categories to arise inductively; Find latent categories; Find patterns across Text; Handle large and diverse corpora; Find key differences between categories

**DO NOT**: Find the “one” best way to categorize text; Capture the categories you want; Tell you who does what to whom; Magically reveal meaning.

4/ LDA

# Study cases

1/ Image of a company during a conference.

2/ Terrorism (enrolment, attack probability, response to general population).

3/ Scientific worldwide success: CERN (Higgs), Curiosity, Philae, Neural net (image recognition, car driving), Theranos, XNA[2](#_ENREF_2), New DNA base pair[3](#_ENREF_3),

4/ Correlation between kid literature and society core values.

# board

**On board:**

**Declined:**

**To be contacted:**

**Susan Marqusee:** Director of the California Institute for Quantitative Biosciences (QB3), and Professor of Molecular and Cell Biology at University of California, Berkeley. Education Director of Synberc, a synthetic biology consortium of UC Berkeley, UC San Francisco, Stanford, Harvard, and MIT. One of the 100 Most Influential Women in the Bay Area (San Francisco Business Times).

**Andrew Ng, David Blei, Micheal Jordan**: LDA fathers.

**Peter Hayes:** Executive Director of the Nautilus Institute for Security and Sustainability.

One lawyer. One sociologist. One economist. One Data Science company CEO.

Notes and references

1 When a company agrees to give SMART data to be analyzed, the data become the SMART property. SMART guarantees the confidentiality of the data (stored on a private server). Data cannot be suppressed due to the fact that we are using a non-relational approach so that the same information could be shared by different sources and therefore, refereeing to the confidentiality agreement, no data can be suppressed. Asking if one information is shared by other sources is a breaking of the confidentiality agreement.

2 Pinheiro, V. B. *et al.* Synthetic Genetic Polymers Capable of Heredity and Evolution. *Science* **336**, 341-344, doi:10.1126/science.1217622 (2012).

3 Malyshev, D. A. *et al.* A semi-synthetic organism with an expanded genetic alphabet. *Nature* **509**, 385-388, doi:10.1038/nature13314 (2014).