PROPOSAL PROJECTS - NLP/IR COURSE

Contact person: Gianni Barlacchi (barlacchi@fbk.eu)

Evaluation and parameters tuning of a search engine

Difficulty: *Low*

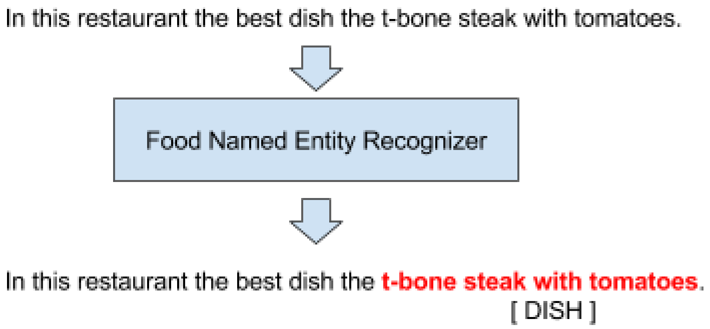
In many ways, search engines have become the most important tool for our information seeking. Due to their tremendous economic value, search engine companies constantly put major efforts to improve their search results. Measuring search effectiveness is thus an important issue because it gives a numeric idea of how good is your system.
We have an huge dataset with real search engine data, with query plus metadata (e.g. information about time and click). The goal of this project is to evaluate the system in terms of relevance of results to user's information need. Then, try to improve the performance find a better parameters setting.
Some useful resources:

- https://www.youtube.com/watch?v=ds1OKuB7IDw
- http://web.simmons.edu/~benoit/lis466/Evaluation-chap8.pdf

Web Page classification and Named Entity Recognition in the food domain

Difficulty: *Medium*

Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names. However, the labels depend from the domain, and thus the training is also domain dependent.
Given a set of restaurant web sites, the challenge of this project is to firstly, identify the menu web page using machine learning classification techniques, and secondly, build a model, specific for the food domain, to recognize and label the dishes present in the page. Preferably, the model has to built for italian language.



In this restaurant the best dish the t-bone steak with tomatoes.

In this restaurant the best dish the **t-bone steak with tomatoes**.

Some useful resources:

- http://open.blogs.nytimes.com/2015/04/09/extracting-structured-data-from-recipes-using-conditional-random-fields/?_r=0
- http://nlp.stanford.edu/software/CRF-NER.shtml

Understanding the Italian food

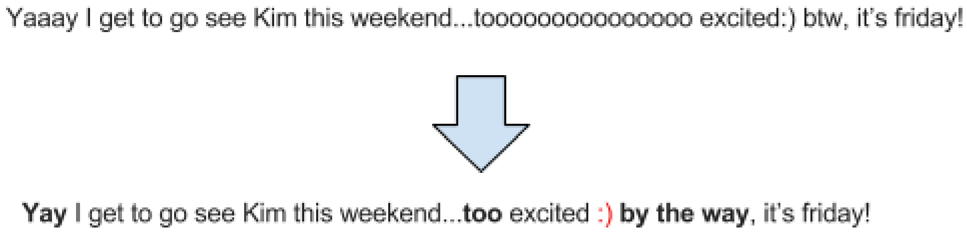
Difficulty: *High*

This project is a mixture between natural language processing and data science.
The goal is to produce a significant statistical analysis of the food habits in Italy. Using a real dataset containing thousand of dishes and recipes in italian language, the idea is to do a study similar the one done by the famous linguist Dan Jurafsky with american menus. He analyzed the language of food highlighting difference between restaurants, cuisine and countries. In this project we would like to do a similar study over the italian territory, understanding for example how food changes between north and south of Italy and between big and small cities.

Social Media text preprocessing

Difficulty: *Medium*

In this project the student has to investigate how to manage row social media texts (e.g Facebook post).
These text usually include many not common words or symbols such as slang, abbreviation and emoticons. Applying a combination of rule based and machine learning approaches, the goal of this project is to build an algorithm that, given in input a social media text, produce in output a clean and analysable version of it, preferably, annotating also smiles in the text.



Yaaay I get to go see Kim this weekend...tooooooooooooooo excited:) btw, it's friday!

Yay I get to go see Kim this weekend...**too** excited :) **by the way**, it's friday!

Embedding methods for the food domain

Difficulty: *High*

The goal of this project is to exploit the joint use of text embedding techniques (e.g word2vec) with network embedding techniques (e.g. LINE) in order to create a similarity measure for object in the food domain (e.g. dishes). Given a list of ingredients and an ingredients network, the idea is to embed this information in order to create an "embedding" version of the dish that is considering the text, but also the network.
At the end, there will be the possibility to evaluate the goodness of this measure with real world search engine data.

References:

- http://arxiv.org/abs/1503.03578
- http://www.thespermwhale.com/jaseweston/papers/fbqa.pdf
- http://arxiv.org/abs/1111.3919