

Predicting emoticons in text

Weekly plan

By Gornishka, Loor, Qodari, Xumara

November 6, 2014

Week 45 (current week)

Priority:

Get a clear idea about the project:

1. how much data do we need - start from around 1000-2000 training tweets and increasing the number in case it appears both possible and needed.
2. how should we pre-process the data - remove hyperlinks and tags; detect the language and make sure it's English; perform spelling and grammar correction; store hashtags separately - they might be useful as features (e.g. tweet containing the hashtag #BestDayEver is more likely to also contain ":)") instead of ":(")
3. decide on "emoticon classes" - start with **Positive** (:) , ;) , :D, :]), **Negative** (:(, ;(, :[) and **Neutral** (:— , :/) and increase the number of classes once the project has progressed enough (possibly extremely positive (:D), extremely negative (;) , surprised (:O) or other, depending on the temporal results).
4. which algorithms are we going to use for classification - possibly **CRF**, **logistic regression** or **SVM**. Implement at least one and use at least one more in order to compare results.
5. what features might be useful - word count; count of positive and negative words in the tweet; count of positive and negative words in the hashtags; special punctuation used (?, !); upper-case words; count of adjectives or other parts of the speech; Expand the list as much as possible through the whole project.

In order to be sure about the points above, research previous work related to the project during week 45 and week 46.

Week 46:

Priorities:

1. Finalize the project idea.
2. Effort on collecting and pre-processing the data.
3. Implement basic functions for feature-extraction.

Week 47:

Priorities:

1. Implement a classifier and possibly get a first working version of the project.
2. Discuss the progress so far. Decide if more data is needed, if the chosen classifiers yield a good performance and what features seem to be important so far.

Week 48:

Priorities:

1. Explore more features and adjust parameters in order to improve the performance.

Week 49:

Priorities:

1. Prepare a demo.
2. Fix possible bugs.
3. Collect results for the presentation and the final report.

Week 50:

Priorities:

1. Prepare a presentation.

Week 51:

Priorities:

1. Finalize the project.
2. Prepare final report.