Unicamp, 12 de setembro de 2017

Fabio M. Tanada RA 880.315

**Relatório IA389 – T2**

# Análise de Sentimentos em Textos

## Objetivo

O objetivo desta tarefa é expor os alunos aos desafios práticos envolvidos na análise de textos e na atribuição de valores de valência ou rótulos de emoções a sentenças.

Não é objetivo desta tarefa avaliar a acurácia de detecção ou a eficiência do modelo implementado mas a análise crítica do projeto e o amadurecimento em relação ao problema.

## Descrição da Tarefa

Esta tarefa deverá ser realizada individualmente ou em dupla.

Deve-se escolher um entre os dois problemas propostos abaixo.

### Problema 1 - Determinação de Valência em Manchetes de Jornais Brasileiros no 1° Semestre de 2017

* Serão fornecidas 500 manchetes dos jornais “Valor Ecônomico”, “O Globo”, “Folha de São Paulo”, “O Estado de São Paulo”, extraídas em datas específicas dos meses de dezembro de 2016 a agosto de 2017. O acesso à base deve ser realizado pelo link: <https://github.com/pdpcosta/manchetesBrasildatabase>
* O problema consiste em classificar as manchetes em valência positiva, neutra ou negativa, associado ao indicador de intensidade. Exemplo: numa escala de -100 a 100, uma manchete classificada como +80 pode ser considerada muito positiva e uma manchete classificada como 5 pode ser considerada neutra.
* A tarefa consiste em definir a abordagem ao problema, o modelo de classificação, as regras de análise e deverá realizar uma implementação prática do algoritmo definido.

Relatório da Tarefa

Esta tarefa apresentou dois desafios claros para mim: o primeiro foi o aprendizado da linguagem Python e o segundo aprender esta nova área da computação.

Para o primeiro desafio eu iniciei curso no coursera do professor Fábio Kon da USP (curso de 9 semanas, estou atualmente na semana 4), além de muitas consultas a internet para tópicos específicos da linguagem, como no stackoverflow.com e docs.python.org. Minha experiência anterior tem ajudado, tenho muita fluência em programação orientada a objeto com C++ e recentemente aprendi também Java e C#.

No segundo desafio realizei pesquisa sobre o tema de “Análise de Sentimento”, tanto no intuito de encontrar referências sobre modelos mais usados quanto exemplos de dicionários léxicos e mesmo de códigos / bibliotecas Python. Abaixo resumo algumas das fontes que consultei:

1. Modelo de análise de sentimento – usei referências como Eduardo Duarte [2] para definir um layout inicial do algoritmo de análise.
2. Dicionários léxicos: Floresta (NLTK), SentiLex-PT [3] e Cláudia Freitas [4].
3. Ferramentas: avaliei monkeylearn.com, mas sem identificar como poderia usar esta ferramenta para este trabalho,
4. Código Python: com certeza a referência mais comum é a da NLTK (Natural Language Tool Kit).

Com base neste estudo inicial desenvolvi a seguinte estratégia:

1. Desenvolver código para ler o arquivo csv e jogar os dados em uma matriz (vetor de vetor) – usei csv.DictReader;
2. Usar código de quebra de sentenças em palavras, do NLTK: nltk.word\_tokenize
3. Usar código para descartar stopwords: nltk.corpus.stopwords.words
4. Usar código que associa classe gramatical às palavras: floresta.tagged\_words()
5. Com base na lista de *tagged words* de cada manchete, busquei avaliar os adjetivos quanto ao sentimento, usando o dicionário SentiLex-PT.
6. Por fim, para uma amostra de 50 manchetes avaliei o resultado deste algoritmo contra uma avaliação pessoal do sentimento dessas manchetes

Como era de se esperar, o primeiro resultado não foi muito animador – neste universo de manchetes de jornais são poucas as que usam adjetivos e menos ainda as que tem adjetivos com uma valência clara (positiva ou negativa). Isto me forçou a procurar novos dicionários, agora para aplicar contra substantivos e verbos.

Após nova busca encontrei o trabalho de Claudia Freitas. Modifiquei o código para acrescentar o teste de substantivos e verbos nos casos em que não houvesse adjetivo na manchete E no caso de verbos usei o stemmer (stemmer.stem) para usar apenas o radical do mesmo na comparação.

Nova rodada de testes e verifiquei que os resultados melhoraram, mas o julgamento de algumas manchetes ainda era incorreto, basicamente devido à limitação do léxico de Claudia Freitas. Acrescentei alguns elementos novos por mim mesmo e os resultados acredito que chegaram a um ponto aceitável, apesar de ainda haver alguns julgamentos incorretos. O meu programa apresenta apenas 3 níveis de valência, positivo – neutro – negativo (valores 1, 0, -1).

Programa

No diretório temos os seguintes arquivos:

* “Sentiment Analysis – FMT – 20170907.py” – programa em Python que implementa o algoritmo descrito na sessão anterior.
* “SentiLex-flex.csv“ – arquivo com o dicionário léxico que avalia sentimento de adjetivos
* “LexicoClaudiaFreitas-n.csv” – arquivo com dicionário de substantivos, baseado no trabalho de Claudia Freitas mas estendido com novas palavras por mim.
* “LexicoClaudiaFreitas-v.csv” – arquivo com dicionário de verbos, baseado no trabalho de Claudia Freitas mas estendido com novas palavras por mim.
* “manchetesBrasildatabase.csv” – arquivo de entrada com as manchetes a serem classificadas.
* “manchetesBrasildatabase - classified.csv” – arquivo de saída com as manchetes já classificadas para valência (valores -1, 0 e 1).

Observação: o arquivo de entrada original apresentava 3 linhas incompletas (sem todos os campos) – detectei manualmente e removi.

Próximos Passos

Para melhorar mais os resultados acredito que teria que mudar a abordagem de análise por palavras para passar a *parser trees* e lidar com n-grams, tentando combinar palavras e lidar principalmente com negativas.

Um último passo imaginei combinar com *machine learning*, ou seja, além da capacidade de *parsing* e análise ter experiência de resultados similares anteriores.

Bibliografia

* [1] “Examples of Portuguese Processing”, <http://www.nltk.org/howto/portuguese_en.html> and <http://www.dainf.ct.utfpr.edu.br/~kaestner/Konstanz/Examples%20for%20Portuguese%20Processing.pdf>
* [2] Duarte, Eduardo Santos. “Sentiment Analysis on Twitter for thePortuguese Language”. Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Dezembro de 2013.
* [3] Carvalho Paula; Silva, Mário. “SentiLex-PT: Principais Características e Potenciais”. Oslo Studies in Language 7, p. 425-438, 2015. https://www.journals.uio.no/index.php/osla/article/download/1444/1341
* [4] Freitas, Cláudia. “Sobre a construção de um léxico da afetividade para o processamento computacional do português”, RBLA Belo Horizonte, v. 13, n. 4, p. 1031-1059, 2013