CAMPUS
DE EXCELENCIA
INTERNACIONAL

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS

MÁSTER UNIVERSITARIO EN INTELIGENCIA ARTIFICIAL

TRABAJO DE FIN DE MÁSTER:

TWEETSC: CORRECTOR DE TEXTO PARA TWITTER

JAVIER MORENO VEGA

TUTOR DE PROYECTO:
OSCAR CORCHO GARCÍA

CO-TUTOR DE PROYECTO:
VÍCTOR RODRÍGUEZ DONCEL

`http://tweetsc.github.io`

11 de mayo de 2018

Índice

1. Introducción

Prueba referencia [1]

1.1. Motivación

1.2. Objetivos

1.3. Resumen del documento

2. Estado del arte

En la actualidad, la normalización lingüística de tuits [8] supone un campo de gran interés y en donde la mayoría de trabajos se han realizado sobre textos en inglés y pocos en español. Además no hay ningún trabajo en donde se incluya, dentro de la normalización de tuits, el estudio de los hashtags o etiquetas y los emoticonos, y su contexto. Una introducción al tema de normalización de tuits es el artículo [1], donde se revisa el estado del arte en NLP sobre variantes SMS y tuit, y cómo la comunidad científica ha respondido por dos caminos: normalización y adaptación de herramientas. El artículo [8] es una buena referencia en el campo de la normalización de tuits en inglés. En donde para detectar palabras fuera de diccionario (OOV) utilizan GNU aspell, y los usuarios (@usuario), los hashtags y las URLs son excluidas de la normalización. En adaptación de herramientas es interesante el trabajo [9] que replantea el tema de reconocimiento de entidades nombradas en corpus de tuits. Combina un clasificador KNN con CRF (Conditional Random Fields).

Una introducción a la normalización de tuits en español es [2][6]. Utiliza la herramienta Freeling [10] para detectar palabras OOV. Uno de los sistemas de normalización de tuits en español, que participó en Tweet-Norm 2013 [2], es [3], que usa reglas de preproceso, un modelo de distancias de edición adecuado al dominio y modelos de lengua para seleccionar candidatos de corrección según el contexto. El sistema obtuvo resultados superiores a la media en la tarea [2][11]. Una mejora a este trabajo por los mismos autores es [4]. En el trabajo [5] hace uso de una combinación de varios "módulos expertos" independientes, cada uno especializado en una tarea concreta de la normalización de tuits, en lugar de centrarse en una sola técnica. En este trabajo

además realiza un estado del arte actual de la normalización de tuits y en concreto para el idioma español.

Un campo muy relacionado con la normalización de tuits es el análisis de sentimientos y un trabajo que realiza un estudio sobre técnicas de análisis de sentimientos de tuits en español es [12]. El trabajo [13] se centra en una técnica Naive-Bayes para el análisis de sentimientos en tuits en español.

Sistemas que participaron en la tarea Tweet-Norm 2013 y que son públicos: Vicomtech [3] [4] [14] RAE (Mejores resultados) [15]

3. Análisis y diseño

3.1. Metodología de desarrollo

3.2. Análisis de requisitos

3.3. Solución propuesta

4. Implementación

4.1. Introducción

4.2. Readme Github

4.3. Javadoc

5. Evaluación

5.1. Metodología

5.2. Corpus

5.3. Goal Standard

5.4. Experimentos

6. Apéndices

6.1. Apéndice A: Bibliografía

Referencias

[1] Wikipedia. Serie temporal — wikipedia, la enciclopedia libre. `https://es.wikipedia.org/w/index.php?title=Serie_temporal&oldid=86168679`, 2015. [Internet; descargado 25-octubre-2016].

[1] What to do about bad language on the internet. Eisenstein, Jacob. 2013. En Proceedings of NAACL-HLT, pp 359–369.

[2] Introducción a la tarea compartida Tweet-Norm 2013: Normalización léxica de tuits en español. Iñaki Alegria, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, Arkaitz Zubiaga. Tweet Normalization Workshop at SEPLN 2013: An overview.

[3] Lexical Normalization of Spanish Tweets with Preprocessing Rules, Domain-Specific Edit Distances, and Language Models. Pablo Ruiz, Montse Cuadros and Thierry Etchegoyhen. Proceedings of the Tweet Normalization Workshop at SEPLN 2013. IV Congreso Español de Informática, Sep 2013, Madrid, Spain.

[4] Lexical normalization of Spanish tweets with rule-based components and language models. Pablo Ruiz, Montse Cuadros and Thierry Etchegoyhen. Procesamiento del Lenguaje Natural. 2014, 52: 45-52.

[5] A modular approach for lexical normalization applied to Spanish tweets. J.M. Cotelo. F.L. Cruz, J.A. Troyano, F.J. Ortega. Expert Systems with Applications, Volume 42, Issue 10, June 2015, pp 4743-4754.

[6] TweetNorm: a benchmark for lexical normalization of Spanish tweets. Iñaki Alegria, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, Arkaitz Zubiaga. Language Resources and Evaluation. December 2015, volume 49, Issue 4, pp 883-905.

[7] A Sentiment Analysis Model of Spanish Tweets. Case Study: Colombia 2014 Presidential Election. Cerón-Guzmán, Jhon Adrián. Maestría thesis, Universidad Nacional de Colombia - Sede Bogotá. 2016.

[8] Lexical normalisation of short text messages: makn sens a twitter. Bo Han, Timothy Baldwin. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, volume 1, pp 368-378. June 2011.

[9] Recognizing named entities in tweets. Xiaohua Liu, Shaodian Zhang, Furu Wei, Ming Zhou. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, volume 1, pp 359-367. June 2011.

[10] Freeling. http://nlp.lsi.upc.edu/freeling/

[11] Tweet-norm. http://komunitatea.elhuyar.eus/tweet-norm/

[12] Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of NLP Techniques. Antonio Fernández Anta, Luis Núñez Chiroque, Philippe Morere, Agustín Santos. Procesamiento del Lenguaje Natural, Revista no 50 marzo de 2013, pp 45-52.

[13] TASS: A Naive-Bayes strategy for sentiment analysis on Spanish tweets. Pablo Gamallo, Marcos García, Santiago Fernández-Lanza. Works-

hop on Sentiment Analysis at SEPLN (TASS2013) (pp. 126-132). 2013.

[14] Sistema Vicomtech. https://github.com/pruizf/tweet-norm-es

[15] Word Normalization in Twitter Using Finite-state Transducers. J Porta, JL Sancho. Tweet-Norm@ SEPLN 1086, 49-53. 2013.

6.2. Apéndice B: Glosario de Términos