Escuela Técnica Superior de Ingenieros Informáticos

Universidad Polítecnica de Madrid

TweetSC: Corrector de texto para twitter

Trabajo Fin de Máster

Máster Universitario en Inteligencia Artificial

AUTOR: Javier Moreno Vega
TUTOR/ES: Óscar Corcho García y
Víctor Rodríguez Doncel

https://jmorenov.github.io/TweetSC/

3 de julio de 2018

RESUMEN

Esta memoria explica todo el trabajo desarrollado entrando en detalle en el estado del arte y el módulo software desarrollado. Primero se realiza una introducción al tema y se exponen los objetivos a realizar. En segundo lugar se presenta el estado del arte sobre el tema de la corrección de textos, específicamente en twitter y en español. Posteriormente explicamos la solución propuesta con todas sus fases. A continuación se presenta la implementación desarrollada y la documentación del código, además de los recursos utilizados. En siguiente lugar evaluamos los resultados. Y por último se desarrollan las conclusiones y líneas futuras.

SUMMARY

This report explains all the work developed by going into detail in the state of the art and the software module developed. First an introduction to the subject is made and the objectives to be made are exposed. Second, the state of the art on the subject of normalization is presented, specifically on twitter and in Spanish. Later we explain the proposed solution with all its phases. Below is the developed implementation and documentation of the code, in addition to the resources used. Next, we evaluate the results. And finally the conclusions and future lines are developed.

Índice

Índice de figuras

Índice de cuadros

1. Introducción

1.1. Motivación

Los nuevos sistemas de comunicación como la mensajería instantánea, chats, redes sociales han generado un uso diferente de los idiomas en estos ámbitos, llamado lenguaje tipo chat [51]. Una de estas redes sociales y en la que este trabajo va a centrarse es Twitter. En esta red social predomina el Uso de emoticonos, repetición de vocales o eliminación de las mismas, uso abusivo de mayúsculas o asusencia, siglas de expresiones populares; lo que dificulta el análisis de los textos. Las ventajas que ofrece esta red social para investigar sobre ella son la cantidad de datos en tiempo real y su fácil acceso.

Uno de los principales problemas a la hora de analizar textos procedentes de las redes sociales son los errores gramaticales que suelen contener, así como la presencia de elementos propios de este tipo de foros que requieren de un procesamiento especial (i.e. hashtags, formas de mencionar a otros usuarios o emoticonos y expresiones habituales en las redes). Además, la limitación en el número de caracteres existente en Twitter la convierte en un caso singular dentro de las redes sociales, ya que los usuarios tienden a adaptar su forma de escribir a dicha limitación, omitiendo palabras y creando abreviaturas que dificultan el uso de herramientas genéricas de procesamiento del lenguaje, especialmente a la hora de realizar tareas como el Análisis de Sentimientos.

Los usuarios en twitter tienden a cometer errores tipográficos, abreviaciones, sustituciones fonéticas y estructuras no gramaticales en los mensajes cortos de texto, causando problemas en las herramientas de análisis. Esto es lo que se consideran palabras mal formadas y la detección de las palabras mal formadas es difícil debido al contexto ruidoso. El objetivo es normalizar estas palabra mal formadas.

A parte de un uso puramente de investigación, este tipo de trabajo también es beneficioso para un estudio de marcas o personas y sobre lo que las persones opinan sobre ello en las redes social, ya que sin el proceso de normalización y análisis de sentimientos estaríamos ante millones de datos que costarían mucho trabajo analizar de una forma automática.

1.2. Objetivos

El objetivo principal de este trabajo es la creación de un corrector que "normalice"tweets en español.
Para cumplir con este objetivo principal se ha dividido en los siguientes subobjetivos.

- Acceso a la API de Twitter para obtener tweets.
- Tokenizar tweets.

- Detectar entre los tokens las palabras fuera del vocabulario (Out-of-Vocabulary, OOV).
- Anotar el tipo de palabras OOV.
- Corregir palabras OOV.

Estos subojetivos se cumplirán con su implementación en un módulo software que además estará disponible en una aplicación web [46].

También ejecutaremos este corrector sobre un corpus de tweets disponible en [69] y compararemos nuestros resultados con los que se consiguieron en [3].

1.3. Resumen del documento

Esta memoria explica todo el trabajo desarrollado entrando en detalle en el estado del arte y el módulo software desarrollado. Primero se realiza una introducción al tema y se exponen los objetivos a realizar. En segundo lugar se presenta el estado del arte sobre el tema de la corrección de textos, específicamente en twitter y en español. Posteriormente explicamos la solución propuesta con todas sus fases. A continuación se presenta la implementación desarrollada y la documentación del código, además de los recursos utilizados. En siguiente lugar evaluamos los resultados. Y por último se desarrollan las conclusiones y líneas futuras.

2. Estado del arte

2.1. Introducción

En la actualidad, la normalización lingüística de tweets [31] 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 [16], donde se revisa el estado del arte en NLP sobre variantes SMS y tweets, y cómo la comunidad científica ha respondido por dos caminos: normalización y adaptación de herramientas.

2.2. Normalización

El modelo de canal ruidoso [64] ha sido tradicionalmente la primera aproximación a la normalización de textos. Supone que el texto mal formado es T y su forma normalizada es S, por lo que hay que encontrar: arg max $P(S|T)$, calculando arg max $P(T|S)P(S)$, P(S) es el modelo del lenguaje y $P(T|S)$ es el modelo de error. [8] caracterizan el modelo de error calculando el producto de operaciones de probabilidad en partes de cadenas de caracteres. [68] mejoraron el modelo incorporando información de la pronunciación. [11] modela el proceso de generación de texto a nivel de palabra para mensajes SMS considerando las abreviaturas grafémicas/fonéticas y los errores tipográficos involuntarios como transiciones de estado ocultas del modelo de Markov (HMM) y emisiones, respectivamente. [12] expandieron el modelo de error introduciendo inferencias de diferentes procesos de formación erróneos, de acuerdo con la distribución de errores muestreada.

Mientras el modelo de canal ruidoso es apropiado para normalización de textos, es difícil aproximar la normalización con exactitud, además estos métodos ignoran el contexto alrededor del OOV, el cual ayuda a resolver ambigüedades. La traducción automática estadística (SMT) se ha propuesto como un medio de normalización de texto sensible al contexto, al tratar el texto mal formado como el idioma de origen, y la forma estándar como el idioma de destino. Por ejemplo [5]. Normalización de textos como un problema de reconocimiento de voz [38]. [6] métodos de estado finitos combinando las ventajas de SMS y el modelo de canal ruidoso. [37] usan un enfoque de traducción automática con un preprocesador para la normalización sintáctica (en lugar de léxica).

El problema de estos trabajos anteriores es que requieren datos de entrenamiento anotados a gran escala, lo que limita su adaptabilidad a nuevos dominios o idiomas, mientras que los trabajos [73] y [31], no. Estos trabajos son una buena referencia en el campo de la normalización de tuits en inglés de forma no supervisada. 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. La normalización tiene relación con los correctores de texto [55] pero difiere en que las palabras mal formadas en los mensajes de texto suelen ser intencionadas, para ahorrar caracteres, como identidad social, o debido a la convención en este subgénero de texto. La detección de las palabras mal formadas es difícil debido al contexto ruidoso. El objetivo es normalizar estas palabra mal formadas, además muchas palabras mal formadas son ambiguas y requieren el contexto para poder normalizarlas.

2.3. Adaptación de herramientas

En vez de adaptar el texto a herramientas de análisis otro de los caminos a seguir es adaptar las herramientas de análisis al texto. Destacan los trabajos de reconocimiento de voz [25] [54], reconocimiento de entidades [18] [58] [40], análisis gramatical [21], modelización de diálogos [57] y resumen automático de textos [65].

El reconocimiento de entidades nombradas (NER) es una tarea de extracción de información que busca localizar y clasificar en categorías predefinidas, como personas, organizaciones, lugares, expresiones de tiempo y cantidades, entidades encontradas en un texto. Las soluciones propuestas para NER suelen recaer en tres categorías: Basado en reglas [39], Basada en aprendizaje automático [20] [66] y Métodos híbridos [35]. Con la disponibilidad de datos anotados, Enron [44] y CoNLL03 [62] se han convertidos en los nuevos métodos dominantes. El estudio actual NER se centra principalmente en textos formales, de hecho, el estado del arte actual (CoNLL03) tiene un éxito del 90.8 % en textos formales y 45.8 % en tweets. En el contexto de los textos en Tweets, existe una dificultad en el reconocimiento de entidades nombradas debido a la falta de información y datos de entrenamiento.

El trabajo en el contexto de los textos de Twitter se puede dividir en tres categorías: NER en tweets, NER en no tweets y aprendizaje semi-supervisado para NER. El trabajo principal de NER sobre tweets es [18], en donde se anotan los tweets y se entrena el modelo con CRF. En cuanto a los trabajos de NER sobre no tweets: [39] utilizan reglas manuales para extraer entidades de tipos predefinidos, [75] utilizan HMM (Hidden Markov Model) mientras que [19] usa CRF. En la tercera categoría, aprendizaje semi-supervisado para NER, se encuentran los trabajos de [36] que utiliza un algoritmo de bootstrapping balanceado, [74] también utiliza un algoritmo de bootstrapping, [43] clusters de palabras, [9] aprende desde texto sin etiquetar y [30] introduce Latent Semantic Association (LSA) para NER. El trabajo más importante y actual de NER para tweets es [40] donde replantea el tema de reconocimiento de entidades nombradas en corpus de tuits. Combina un clasificador KNN con CRF (Conditional Random Fields).

La desambiguación léxica o etiquetado gramatical (POST) es una parte muy importante y útil en la tarea de normalización de textos ya que nos permite definir el

subconjunto de palabras debido a su categoría gramatical que con una probabilidad pueden ser la normalización de un OOV. Además un gran porcentaje de palabras en un texto son palabras que pueden ser asignadas a más de una clase morfológica, a más de un part-of-speech (PoS). Uno de los trabajos más importantes y probado para español es [67], este trabajo presenta un método de POST de ventana deslizante (SWPoST), asigna el part-of-speech de una palabra basado en la información que dan las palabras en una ventana fija de alrededor. Puede ser implementado como una máquina de estados finitos (Máquina de Mealy).

2.4. Normalización en español

Una introducción a la normalización de tuits en español es [3][2]. Este trabajo propuso en 2013 una tarea o competición en la que los participantes proponían soluciones de normalización de tweets. Los organizadores de la competición ofrecían dos datasets de tweets ya notados uno de desarrollo y otro para test, junto con un tercero que no era público y que era usado para la última evaluación.

Las soluciones ofreciddas por los participantes se pueden dividir en dos categorías, los que utilizan generación de candidatos junto un modelo del lenguaje, y los que utilizan transductores o FSTs (Finite State Transducers). El participante que mejor accurancy consiguió, Sistema RAE [23] con un 0.781, optó por la segunda categoría e implementó un siststem basado en FSTs para la tarea de normalización léxica de mensajes de Twitter en Español. El sistema desarrollado consiste en transductores que se aplican a tokens OOV. Los transductores implementan modelos de variación lingüística que generan conjuntos de candidatos acordes a un léxico. Un modelo estadístico del lenguaje se usa para obtener la secuencia de palabras más probable. El sistema tiene tres componentes principales que se aplican secuencialmente. Un analizador que ejecuta tokenización y análisis léxico sobre palabras en forma estándar y otras expresiones (números, fechas, ...). Un componente que genera palabras candidatas para los tokens OOV. Un modelo estadístico del lenguaje para obtener la mejor secuencia de palabras. Y finalmente un truecaser para capitalizar correctamente las palabras asignadas a los tokens OOV. El conjunto de confusión de un token OOV se genera aplicando el algoritmo de camino mínimo a la expresión: $W \circ E \circ L$. Donde W es el automata que representa el token OOV, E es un transductor de editado que genera todas las posibles variaciones de un token, y L es un conjunto de palabras objetivo. Dentro de esta categoría se encuentran los trabajos de la tarea: [1] en donde usan una batería de módulos para generar diferentes propuestas de corrección para cada palabra desconocida. La corrección definitiva se elige por votación ponderada según la precisión de cada módulo, [3] que además utiliza un modelo para el reconocimiento de voz para la generación de candidatos y [34] presentan dos estrategias basadas en FSTs una con reglas diseñadas manualmente y la otra automática.

Entre los participantes que optaron por la primera categoría destaca [59][71] 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. Su arquitectura está formada por: preproceso basado en expresiones regulares y listas customizadas, generación de candidatos mediante una técnica de mínima de distancia de editado, ranking de candidatos mediante una combinación con pesos de la puntuación del modelo del lenguaje y la distancia de editado y la puntuación del modelo de lenguaje es n-grama utilizando la distancia Levenshtein. El sistema obtuvo resultados superiores a la media en la tarea. Una mejora a este trabajo por los mismos autores es [60] en donde utilizan un sistema basado en reglas para seleccionar los candidatos. Otros trabajos en esta categoría son: [24] que propone un sistema basado en [32], [70] y [49] que emplea técnicas de RAH (reconocimiento del habla) mediante la herramienta TENOR [47] junto con un modelo del lenguaje. Otro trabajo basado en la tarea de Tweet-Norm pero que no participó en ella es [10], ellos optaron por normalizar los OOV basándose en similaridad entre grafemas y fonemas; generan el conjunto de confusión (de candidatos) usando grafemas y fonemas, seguido de transductores aplicados mediante reglas para las palabras extranjeras y acentos, la selección de candidatos mediante un modelo del lenguaje con la herramienta Kenlm [33].

Fuera de estas dos categorías nos encontramos con los trabajos: [61] que utiliza conversiones basadas en reglas hasta una forma final normalizada. Después de recibir una lista con las posibles correcciones el sistema selecciona la más común acorde con una lista de palabras ordenada por frecuencia, [72] utilizan una lista de prioridad para los candidatos obtenidos y una tabla de frencuencias de palabrar para puntuarlos, [32] presentan una estrategia basada en búsquedas rápidas mediante una lista de frecuencias aprendida desde un corpus de tweets, [76] no generan candidatos simplemente selecciona palabras OOV y las corrigen con un corrector externo y [50] generan candidatos y seleccionan el mejor mediante una función de distancia. Una mejora a este último trabajo por parte de los autores fue [13] donde añaden un modulo de puntuación para la selección de candidatos.

Otros trabajos sobre normalización en español son [47] en donde se generan candidatos con indexación fonética y se seleccionan el candidato calculando la similaridad léxica junto con un modelo del lenguaje trigrama y [53]. Estos trabajos son principalmente sobre mensajes SMS, y no abordan la normalización de tuits en su conjunto. Dentro de la normalización en español existen otras tareas relacionadas como es la tokenización y aquí destaca el trabajo [26] que estudia la tokenización de textos SMS.

2.5. Word2Vec

Muchos sistemas y técnicas actuales de NLP tratan las palabras como unidades atómicas, no hay noción de similaridad entre palabras y son representadas como índices en un vocabulario, por ejemplo el modelo N-grama, para tratar de resolver

este problema aparecen las representaciones continuas de palabras. Las representaciones continuas de palabras entrenadas sobre corpus sin etiquetas son útiles para muchos trabajos de NLP. Muchos tipos diferentes de modelos han sido propuestos para estimar representaciones continuas de palabras, incluyendo Latent Semantic Analysis (LSA) y Latent Dirichlet Allocation (LDA). En este trabajo se centran en las representaciones distribuidas de palabras aprendidas por redes neuronales, ya que se demostró que su eficacia era considerablemente mejor que LSA para preservar regularidades lineales entre palabras, LDA además es computacionalmente caro en datasets grandes. Además se ha demostrado las redes neuronales basadas en modelos del lenguaje mejoran significativamente los modelos N-grama [7] [41] [63].

Los dos modelos de redes neuronales que destacan basados en modelos del lenguaje son: Feedforward Neural Net Language Model (NNLM) [7] y Recurrent Neural Net Language Model (RNNLM) que mejora algunas limitaciones de NNLM. El problema de estos modelos es que con grandes cantidades de datos son muy costotos computacionalmente. Para resolver este problema en el trabajo [42] desarrollado por Google se presentaron dos nuevos modelos de arquitecturas para calcular representaciones continuas de vectores de palabras a partir de grandes datasets, además crearon un framework de bibliotecas llamado Word2Vec [28]. El principal objetivo de este trabajo es introducir técnicas que puedan ser usadas para aprender vectores de palabras de gran calidad a partir de grandes datasets con millones de palabras y con millones de palabras en el vocabulario. Decidieron explorar modelos más simples que aunque no puedan representar los datos de forma tan precisa como las redes neuronales pero pueden ser entrenados con muchos más datos de forma más eficiente. Estos modelos son: Continuous Baf-of-Words model (CBOW) similar a NNLM, la capa oculta no-lineal se elimina y la capa de proyección es compartida por todas las palabras; y Continuous Skip-gram model similar a CBOW pero en vez de predecir la palabra actual basándose en el contexto intenta maximizar la clasificación de la palabra basándose en otra palabra de la misma frase.

La mayoría de las técnicas de representación continua de vectores de palabras representan cada palabra del vocabulario como un vector distinto, sin parámetros compartidos. En particular se ignora la estructura interna de las palabras lo que es una importante limitación en lenguajes ricos morfológicamente. Para intentar resolver este problema en el trabajo [56], desarrollado por Facebook [17] y llamado fastText, se propone un nuevo enfoque basado en el modelo skipgram [42] donde cada palabra se representa como una bolsa de caracteres n-gramas. Una representación de vector está asociada con cada caracter n-grama, las palabras se representan como la suma de estas representaciones. Al usar una representación de vector distinta para cada palabra, el modelo skipgram ignora la estructura interna de las palabras y en este nuevo trabajo se implementa una función de puntuación diferente para tener en cuenta esta información.

3. Solución propuesta

La solución propuesta y a la que hemos llamado TweetSC (Tweet Spell Checker) [46] se llegó a ella a partir de varios análisis y evaluaciones que se hicieron con diversas bibliotecas y algoritmos, y todos ellos se pueden encontrar en la solución final para su uso.

En la primera versión de nuestra solución se construyó un corrector de texto sencillo basándonos en el creado por Peter Norvig [52], el cuál utiliza un diccionario para seleccionar las palabras incorrectas y las corrige mediante el teorema de Bayes usando probabilidades. Se usa la fórmula: $argmax_{c \in candidates} P(c|w)$, que mediante el teorema de Bayes es equivalente a: $argmax_{c \in candidates} P(c)P(w|c)/P(w)$, y como P(w) es igual para cada candidato c: $argmax_{c \in candidates} P(c)P(w|c)$. Esta fórmula trata de seleccionar el candidato de probabilidad máxima para cada palabra. Para calcular la probabilidad se usan dos diccionarios, uno de palabras en Español y otro de nombres propios. Se utilizó esta primera versión como punto de partida para ir creando versiones más avanzadas.

El resultado final y por tanto nuestra versión definitiva consiste en un proceso iterativo sobre el tweet que se puede dividir en 6 fases: Tokenización, reglas de preproceso, detección de OOVs, generación de candidatos para cada OOV, ranking de candidatos y postproceso.

Además para convertir el sistema en uno más dinámico se ha desarrollado una aplicación web con acceso a la API de Twitter para obtener los tweets mediante querys introducidas en un formulario de nuestra aplicación web.

3.1. Tokenización

Cómo realizan los analizadores léxicos en los compiladores, en la primera fase de nuestro proceso se realiza una tokenización del texto o tweet, un tokenizador genera una salida compuesta de tokens o símbolos.

Para mejorar la versión inicial se hizo uso de la biblioteca Stanford NLP [29]. Esta biblioteca creada por Stanford NLP Group ofrece tanto etiquetado gramatical (POS Tagging o POST) cómo deteccion de etiquetas (Named Entity Recognition o NER), nosotros la hemos utilizado para esta fase de tokenización. Además de StanfordNLP para la tokenización hemos utilizado Freeling [22], también se ha añadido al sistema el análisis que ofrece freeling.

En esta primera fase se recibe como entrada el texto del tweet y genera una lista de tokens que pasaran a la siguiente fase.

3.2. Reglas de preprocesado

Una vez que hemos obtenido todos los tokens de un tweet se aplican unas reglas de preproceso para normalizar palabras típicas de la red social, pictogramas, fonogramas, onomatopeyas, números, acrónimos, etc. Tras aplicar estas reglas a los tokens que las acepten, se crean OOVs con estos token, se anotan como variaciones y se eliminan de la lista de tokens para las fases siguientes. Los OOV generados se añaden a la lista final de OOV.

3.3. Detección de OOV

Esta fase tiene como elementos de entrada los tokens restantes de la fase anterior, y se ejecuta token por token el detector de OOV. Para detectarlos se aplican reglas y se van descartando los tokens que son URLs, usuarios de twitter, hashtag de twitter y fechas; los elementos restantes se comparan con tres diccionarios utilizados como recursos: diccionario de español, diccionario de inglés y diccionario de entidades.

Los token que se detecten dentro del diccionario de español se descartan como OOV, los que se detecten en el diccionario de inglés se anotan como NoEs (No español o ininteligible) y los que se detecten en el diccionario de entidades se anotan como Correct (palabras correspondientes a una entidad o un nuevo préstamo). Para el resto de token que no han sido aceptados en ninguna regla se crea una lista de OOV y son los que pasaran a la siguiente fase pudiendo al final ser anotados como Variation o NoEs.

3.4. Generación de candidatos OOV

La generación de candidatos se puede considerar la primera fase de la corrección en sí, ya que sólo se trabaja con OOV a los que se va a buscar una corrección, en ese caso candidatos para ese OOV. Esta fase tiene como entrada la lista de OOV que no han sido etiquetados en la fase anterior, es decir, los que pueden ser Variation o NoEs. Para cada OOV se generarán una lista de candidatos con diferentes métodos. Los métodos que hemos utilizado con los nombres que hemos definido son: LevenshteinFST, Metaphone, L_L, FastTest.

- **LevenshteinFST:** Método que utiliza un FST (Finite State Transducers) para generar variaciones en el OOV con un máximo de distancia de editado según la distancia Levenshtein.
- **Metaphone:** Método que utiliza el algoritmo del metáfono en español [48]. Su funcionamiento consiste en generar los fonemas de todos los diccionarios que hemos utilizado para después comparar el fonema del OOV y seleccionar los de mayor similaridad con los fonemas de los diccionarios.
- **L_L:** Los candidatos generados con este método son las palabras aceptadas por el lenguaje L(_L)+.

- **FastText:** Este método hace uso de la biblioteca fastText [17], a partir de un modelo generado mediante redes neuronales y representando las palabras de forma continua. Los OOV se convierten a vectores de palabras y se comparan con los vectores del modelo generado para obtener los candidatos más parecidos a partir del la similaridad del coseno.

Estos métodos se ejecutan sobre todos los OOV y generan una lista de candidatos que pasarán a la siguiente fase.

3.5. Ranking de candidatos

El ranking de candidatos es la fase que define la corrección de un OOV, o si no tiene corrección (se anota como NoEs). Para generar este ranking hemos utilizado dos marcadores, uno un modelo del lenguaje N-Gram (Modelo del lenguaje 11) mediante la biblioteca OpenNLP [4] y el otro la distancia de editado Damerau-Levenshtein.

El marcador N-Gram se realiza mediante la comparación de los candidatos de su puntuación en el modelo del lenguaje, es decir, para cada candidato se calcula su puntuación si fuera el elegido. Para el marcador de la distancia Damerau-Levenshtein se calculan la distancia entre el OOV y cada candidato. Finalmente mediante estas dos puntuaciones se calcula se realiza el ranking de candidatos posicionando primero los candidatos con mejor puntuación en el modelo del lenguaje y menor distancia de editado al OOV.

Se ha definido además un umbral mínimo para realizar el ranking, los candidatos que no lo cumplan con los marcadores son eliminados. Al finalizar este proceso para cada OOV se selecciona el mejor candidato del ranking, y se anota como Variation, y si no tuviera candidatos, debido al umbral, se anotan como NoEs.

3.6. Postproceso

Esta última fase consiste en poner mayúsculas en las palabras que fueran necesarias, así como signos de exclamación e interrogación.

4. Implementación

En esta sección se pretende explicar toda la implementación software que se ha realizado de nuestra solución. Primero se realizará una introducción comentando lenguajes y herramientas utilizadas. Segundo se explicará dónde encontrar y cómo utilizar nuestro software . Posteriormente la documentación generada del código fuente. Y por último se explicará la aplicación web que se ha desarrollado.

4.1. Introducción

La implementación se ha realizado en tres módulos o componentes, por una parte tenemos la biblioteca con la funcionalidad necesaria para corregir textos de twitter, acceder a su API y evaluar los resultados sobre un corpus de tweets; después un modulo que implementa aplicación web y por último otro que ofrece funcionalidad para utilizar la biblioteca desde línea de comandos.

El lenguaje principal utilizado en todo el proyecto ha sido Java, con la excepción de Python para los script de evaluación y normalización de archivos de datos, y hemos hecho uso de Google Cloud Engine [27] para que la aplicación web esté disponible para cualquier usuario [45].

Nuestro sistema software se ha intentado desarrollar de forma que sea un sistema de procesamiento dinámico pudiendo añadir y quitar funcionalidad de manera sencilla, cualquier algoritmo o método implementado funciona a partir de una clase superior para que se puedan añadir nuevos métodos.

El módulo principal del sistema es TweetSCCore que implementa la biblioteca para corregir tweets, acceder a la API de twitter y evaluar resultados mediante el script ofrecido por Tweet-Norm 2013 [3]. Los módulos que funcionan a partir de TweetSCCore son: TweetSCExecutable que implementa la funcionalidad necesaria para ejectar nuestro sistema a partir de línea de comandos, y TweetSCWeb donde se implementa la aplicacion web.

Se han implementado dos métodos de corrección o normalización de tweets a los que hemos llamado DictionaryMethod, método preliminar que hace uso de diccionarios y la regla de Bayes, y TweetSCMethod que es nuestro método final con las fases que han sido explicadas.
El diagrama de clases del sistema completo se muestra a continuación:

4.2. Cómo usarlo

Para utilizar nuestro software primero es necesario tener instalado Git y Java 1.8. Después de bajar el código fuente:

```
git clone https://github.com/jmorenov/TweetSC
```

Posteriormente se compila el código:

```
cd TweetSC/code/
chmod +x build_all.sh
./build_all.sh
```

Para ejecutarlo desde línea de comandos:

```
java -jar tweetscexecutable-all-v0.5.0-alpha.jar -text Texto de prueba
```

La ejecución de la evaluación sobre el corpus de Tweet-Norm 2013 [3]

```
java -jar tweetscexecutable-all-v0.5.0-alpha.jar \
    -workingDirectory evaluation \
    -annotatedFile tweet-norm-dev500_annotated.txt \
    -tweetsFile tweet-norm-dev500.txt \
    -resultFile results-test-dev500.txt \
    -method TweetSCMethod
```

La ejecución de la aplicación web:

```
cd tweetscweb
./gradlew run
```

4.3. Aplicación web

En esta sección se explicará el funcionamiento de la aplicación web desarrollada y se mostrarán capturas de pantalla de la misma en funcionamiento.

La aplicación web se encuentra en nuestro paquete TweetSCWeb y hace uso del framework Spring Boot. Para el backend se utiliza Java y para el frontend Javascript (Jquery). Se ha intentado realizar un diseño sencillo y fluido usando la biblioteca Bootstrap.

El funcionamiento de la aplicación web es muy sencillo, tiene una sección principal desde la que se accede a las demás secciones mediante una barra de navegación.

Fig. 1: Inicio de la aplicación web

En la sección siguiente se puede ver un corrector de texto simple.

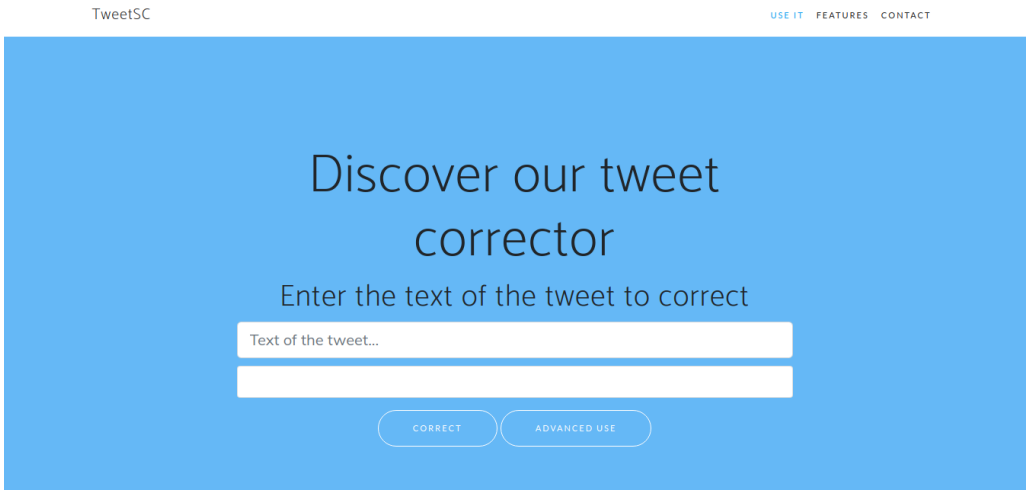

Fig. 2: Sección para utilizar el corrector

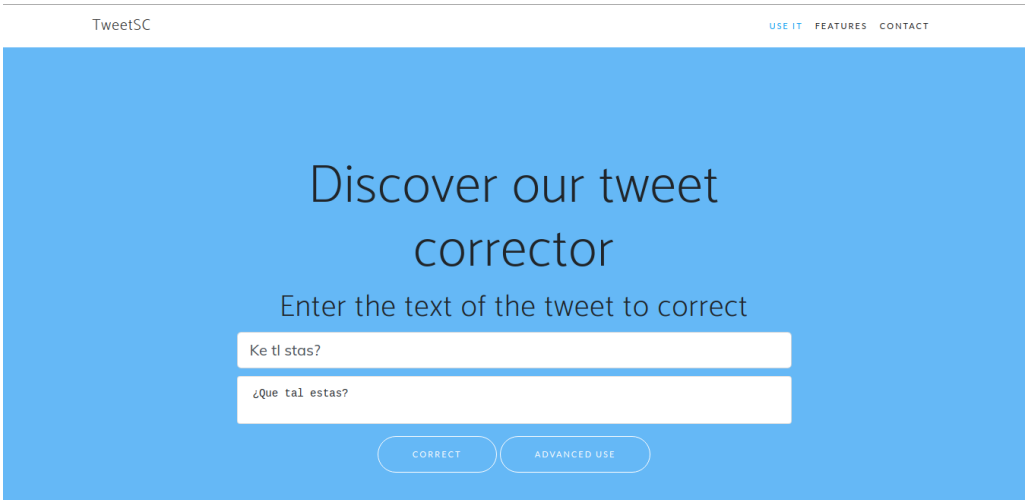

Fig. 3: Ejemplo de texto corregido

Si accedemos al corrector de uso avanzado podemos buscar tweets mediante texto, usuarios o id del tweet.

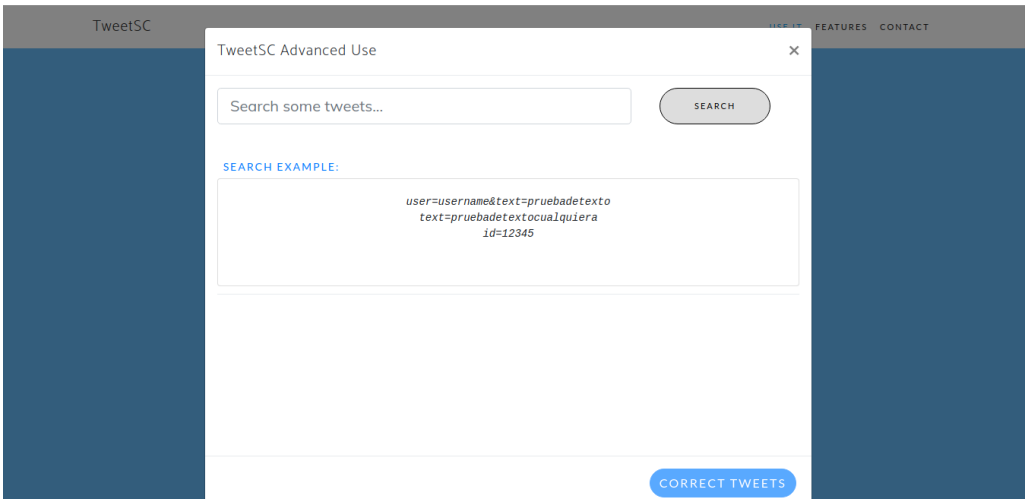

Fig. 4: Corrector de tweets uso avanzado

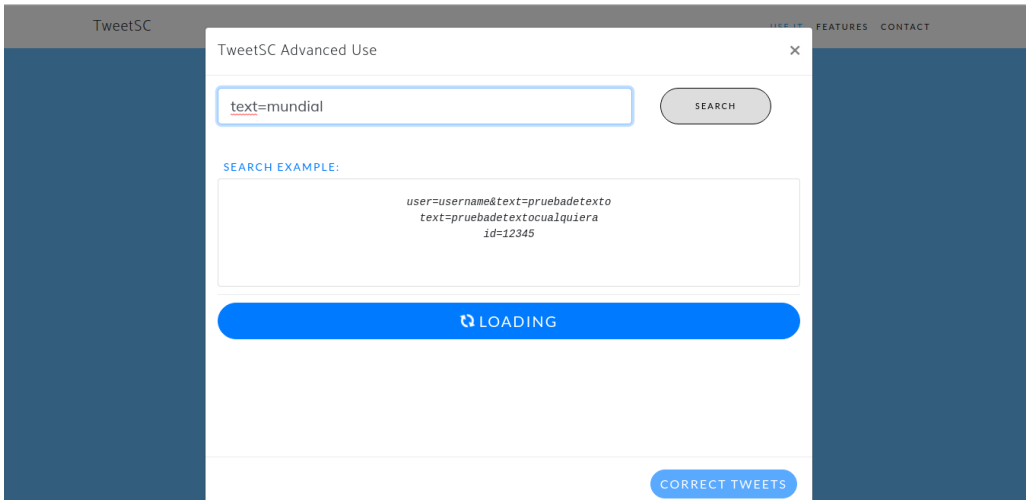

Fig. 5: Ejemplo de búsqueda de tweets

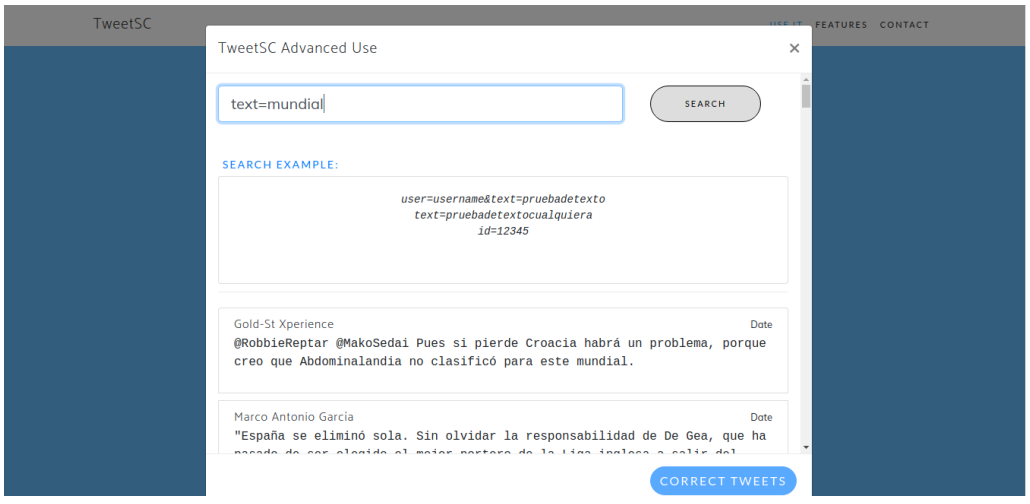

Fig. 6: Ejemplo de tweets encontrados

Los tweets encontrados se pueden seleccionar para corregirlos.

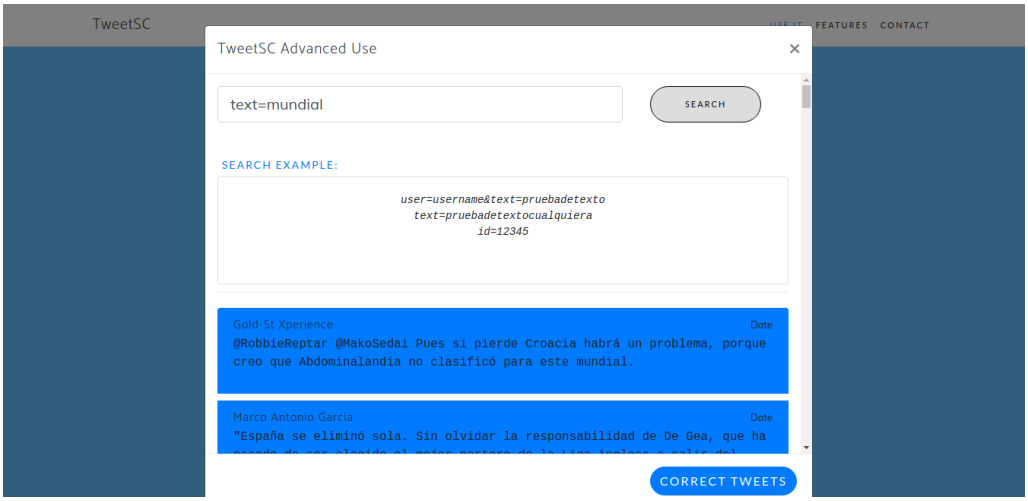

Fig. 7: Ejemplo de selección de tweets para corregir

Cuando los tweets son corregidos se marcan y se muestran ambas versiones, normalizada y la inicial.

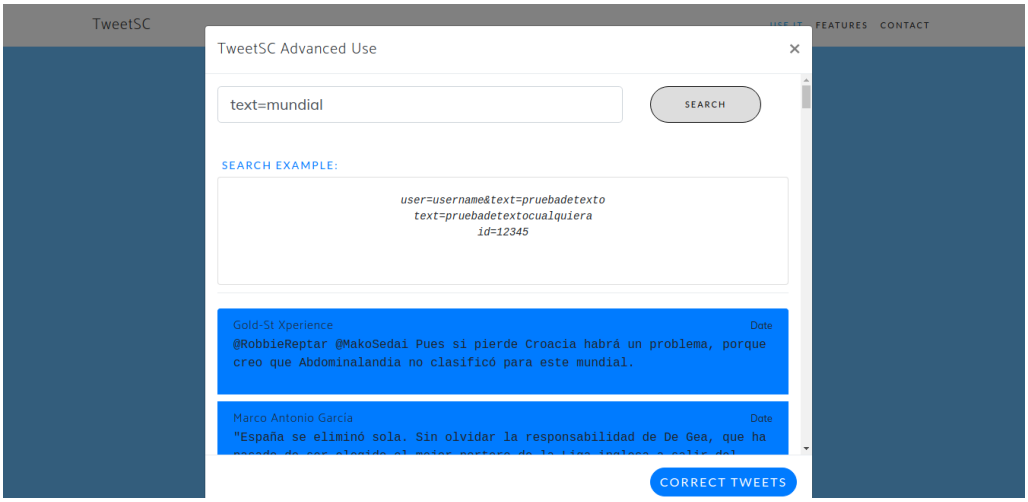

Fig. 8: Ejemplo de tweets corregidos

La siguiente sección es la de características del sistema.

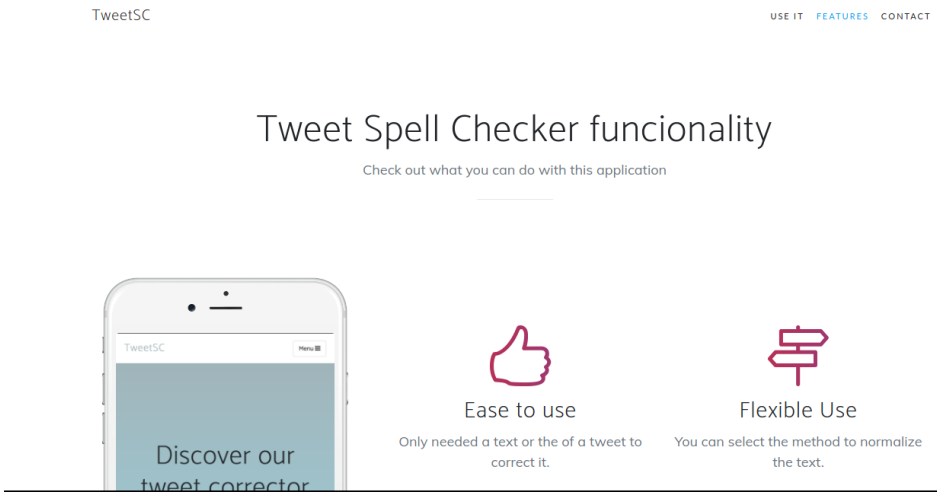

Fig. 9: Sección de características de la aplicación web

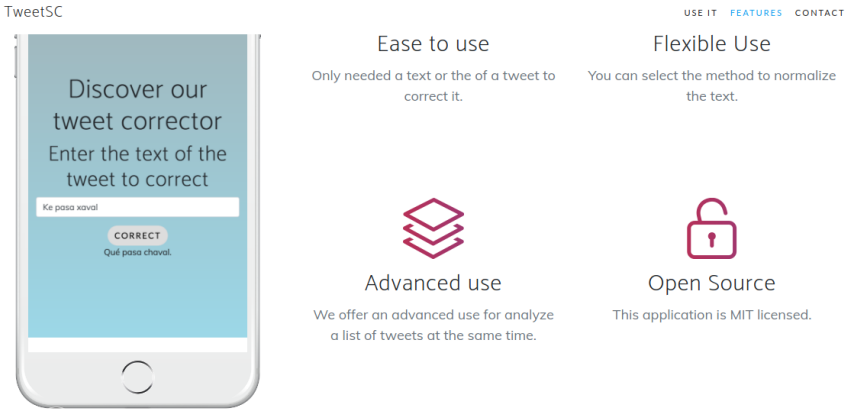

Fig. 10: Sección de características de la aplicación web

Además se ha añadido una sección para colaborar mediante GitHub en el desarrollo.

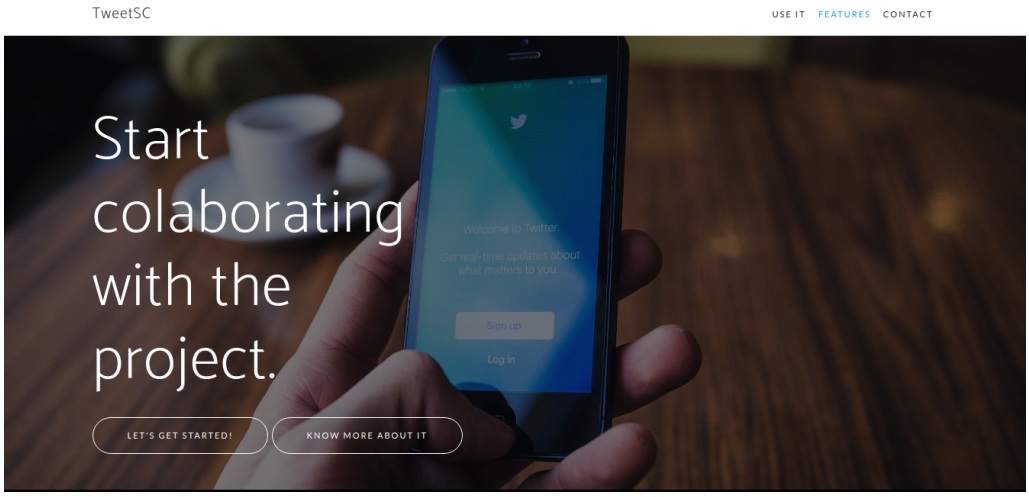

Fig. 11: Sección de colaboración

Por último la sección de contacto.

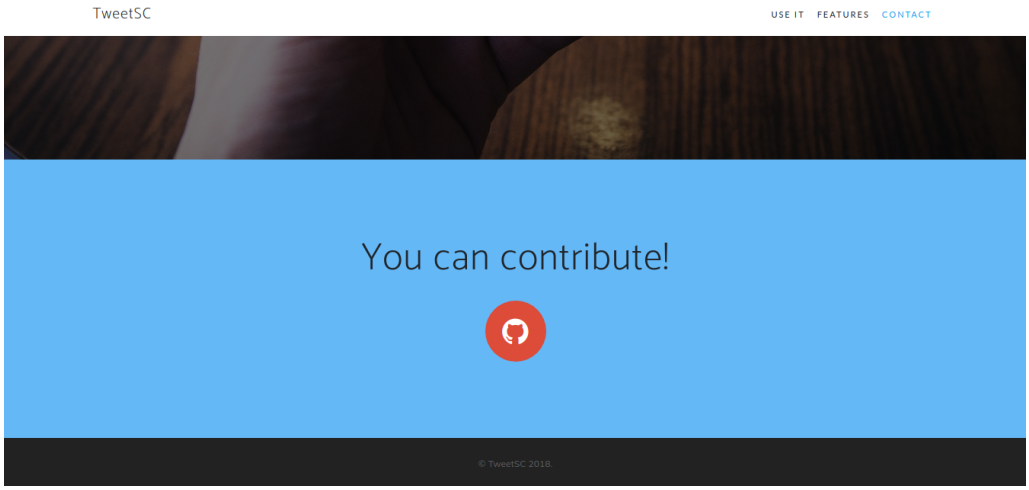

Fig. 12: Sección de contacto

5. TweetSCCore Documentación del código (Java Documentation)

Class Hierarchy

Classes

- java.lang.Object
    - com.jmorenov.tweetsccore.analyzer.AnalysisElement (in 5.5.1, page 39)
    - com.jmorenov.tweetsccore.analyzer.Analyzer (in 5.5.7, page 41)
        - com.jmorenov.tweetsccore.analyzer.FreelingAnalyzer (in 5.5.14, page 42)
    - com.jmorenov.tweetsccore.candidates.Candidate (in 5.6.1, page 43)
    - com.jmorenov.tweetsccore.candidates.CandidatesMethod (in 5.6.7, page 44)
        - com.jmorenov.tweetsccore.candidates.FastTextCandidatesMethod (in 5.6.21, page 47)
        - com.jmorenov.tweetsccore.candidates.LevenshteinFSTCandidatesMethod (in 5.6.28, page 48)
        - com.jmorenov.tweetsccore.candidates.MetaphoneCandidatesMethod (in 5.6.35, page 50)
    - com.jmorenov.tweetsccore.evaluation.TweetNormEvaluationResult (in 5.7.1, page 51)
    - com.jmorenov.tweetsccore.evaluation.TweetNormEvaluator (in 5.7.7, page 52)
    - com.jmorenov.tweetsccore.extra.File (in 5.8.8, page 57)
    - com.jmorenov.tweetsccore.extra.FreelingInitializator (in 5.8.14, page 59)
    - com.jmorenov.tweetsccore.extra.OOV (in 5.8.20, page 60)
    - com.jmorenov.tweetsccore.extra.Parser (in 5.8.26, page 61)
    - com.jmorenov.tweetsccore.method.Method (in 5.9.17, page 67)
        - com.jmorenov.tweetsccore.method.DictionaryMethod (in 5.9.9, page 66)
            - com.jmorenov.tweetsccore.method.DictionaryAnalysisMethod (in 5.9.1, page 65)
    - com.jmorenov.tweetsccore.ner.NER (in 5.4.1, page 36)
        - com.jmorenov.tweetsccore.ner.StanfordNLPNER (in 5.4.14, page 38)
    - com.jmorenov.tweetsccore.ner.NERELement (in 5.4.8, page 36)
    - com.jmorenov.tweetsccore.post.FreeLingPOST (in 5.10.1, page 69)
    - com.jmorenov.tweetsccore.post.OpenNLPPOST (in 5.10.5, page 69)
    - com.jmorenov.tweetsccore.post.POST (in 5.10.9, page 70)
        - com.jmorenov.tweetsccore.post.StanfordNLPPOST (in 5.10.16, page 70)
    - com.jmorenov.tweetsccore.preprocess.ApplyRules (in 5.1.1, page 22)
    - com.jmorenov.tweetsccore.preprocess.Rule (in 5.1.7, page 23)
    - com.jmorenov.tweetsccore.preprocess.Rules (in 5.1.13, page 24)
    - com.jmorenov.tweetsccore.spellchecker.SpellChecker (in 5.11.1, page 71)
    - com.jmorenov.tweetsccore.tokenizer.Tokenizer (in 5.12.29, page 78)
        - com.jmorenov.tweetsccore.tokenizer.FreelingTokenizer (in 5.12.1, page 74)
        - com.jmorenov.tweetsccore.tokenizer.NGramTokenizer (in 5.12.8, page 75)

- com.jmorenov.tweetsccore.tokenizer.OpenNLPTokenizer (in 5.12.15, page 76)
- com.jmorenov.tweetsccore.tokenizer.StanfordNLPTokenizer (in 5.12.22, page 77)

- com.jmorenov.tweetsccore.twitter.Tweet (in 5.2.1, page 26)
  - com.jmorenov.tweetsccore.twitter.TweetCorrected (in 5.2.8, page 29)
- com.jmorenov.tweetsccore.twitter.TwitterConfiguration (in 5.2.15, page 32)
- com.jmorenov.tweetsccore.twitter.api.Search (in 5.3.1, page 33)
- java.lang.Enum
  - com.jmorenov.tweetsccore.candidates.CandidatesMethodType (in 5.6.14, page 46)
  - com.jmorenov.tweetsccore.extra.Annotation (in 5.8.1, page 56)

5.1. Package com.jmorenov.tweetsccore.preprocess

Package Contents *Page*

Classes

5.1.1. Class ApplyRules

ApplyRules class to apply preprocess rules.

5.1.2. Declaration

```
public class ApplyRules
 extends java.lang.Object
```

5.1.3. Constructor summary

ApplyRules() Constructor of the class.

5.1.4. Method summary

apply(String) Method to apply the rules to a text.

5.1.5. Constructors

- **ApplyRules**

```
public ApplyRules() throws java.io.IOException
```

- **Description**
  Constructor of the class.
- **Throws**
  - `java.io.IOException` –

5.1.6. Methods

- **apply**

```
public java.util.List apply(java.lang.String text)
```

- **Description**
  Method to apply the rules to a text.
- **Parameters**
  - `text` – String with the text
- **Returns** – List of OOVs

5.1.7. Class Rule

Rule class that define a rule element.

5.1.8. Declaration

```
public class Rule
 extends java.lang.Object
```

5.1.9. Constructor summary

Rule(String, String) Constructor of the class.

5.1.10. Method summary

getRegex() Method to get the regex of the rule.
getResult() Method to get the result of a rule.

5.1.11. Constructors

- **Rule**

```
public Rule(java.lang.String regex,java.lang.String
    result)
```

- **Description**
  Constructor of the class.
- **Parameters**
  - `regex` – String
  - `result` – String

5.1.12. Methods

- **getRegex**

```
public java.lang.String getRegex()
```

  - **Description**
    Method to get the regex of the rule.
  - **Returns** – String

- **getResult**

```
public java.lang.String getResult()
```

  - **Description**
    Method to get the result of a rule.
  - **Returns** – String

5.1.13. Class Rules

Rule class that define the Rules element.

5.1.14. Declaration

```
public class Rules
 extends java.lang.Object
```

5.1.15. Constructor summary

Rules(String) Constructor of the class.

5.1.16. Method summary

addRule(Rule) Method to add a new rule.
getRules() Method to get the rules.

5.1.17. Constructors

- **Rules**

```
public Rules(java.lang.String rulesFilename) throws java
    .io.IOException
```

  - **Description**
    Constructor of the class.
  - **Parameters**
    - `rulesFilename` – String with the file name of the rules
  - **Throws**
    - `java.io.IOException` – When the file is not found

5.1.18. Methods

- **addRule**

```
public void addRule(Rule rule)
```

  - **Description**
    Method to add a new rule.
  - **Parameters**
    - `rule` – Rule

- **getRules**

```
public java.util.List getRules()
```

  - **Description**
    Method to get the rules.
  - **Returns** – List of Rule

5.2. Package com.jmorenov.tweetsccore.twitter

Package Contents *Page*

Classes

Tweet corrected class with the structure of a corrected tweet.

TwitterConfiguration 32
TwitterConfiguration class with the configuration of the connection with the API of twitter.

5.2.1. Class Tweet

Tweet class with the structure of a tweet.

5.2.2. Declaration

```
public class Tweet
 extends java.lang.Object
```

5.2.3. All known subclasses

TweetCorrected (in 5.2.8, page 29)

5.2.4. Constructor summary

Tweet() Default constructor of the class.
Tweet(Status) Constructor of the class.
Tweet(String, String, String, String) Constructor of the class.
Tweet(String, String, String, String, String) Constructor of the class.
Tweet(Tweet) Copy constructor

5.2.5. Method summary

getDate() Method to get the date of the tweet.
getHash() Method to get the hash of the tweet.
getId() Method to get the id of the tweet.
getText() Method to get the text of the tweet.
getUsername() Method to get the username of the tweet.
toString() Method to get the string of the Tweet.

5.2.6. Constructors

- **Tweet**

```
public Tweet()
```

  - **Description**
    Default constructor of the class.

- **Tweet**

  ```
  public Tweet(Status tweetStatus)
  ```

  - **Description**
    Constructor of the class.
  - **Parameters**
    - `tweetStatus` – Status from the object of Twitter4j.

- **Tweet**

  ```
  public Tweet(java.lang.String id,java.lang.String
    username,java.lang.String hash,java.lang.String text)
  ```

  - **Description**
    Constructor of the class.
  - **Parameters**
    - `id` – String with the id of the tweet.
    - `username` – String with the username of the tweet.
    - `hash` – String with the hash of the tweet.
    - `text` – String with the text of the tweet.

- **Tweet**

  ```
  public Tweet(java.lang.String id,java.lang.String
    username,java.lang.String hash,java.lang.String text,
    java.lang.String date)
  ```

  - **Description**
    Constructor of the class.
  - **Parameters**
    - `id` – String with the id of the tweet.
    - `username` – String with the username of the tweet.
    - `hash` – String with the hash of the tweet.
    - `text` – String with the text of the tweet.
    - `date` – String with the date of the tweet.

- **Tweet**

```
public Tweet(Tweet tweet)
```

- **Description**
  Copy constructor
- **Parameters**
  - `tweet` – Tweet to copy from.

5.2.7. Methods

- **getDate**

```
public java.lang.String getDate()
```

  - **Description**
    Method to get the date of the tweet.
  - **Returns** – String with the date of the tweet.

- **getHash**

```
public java.lang.String getHash()
```

  - **Description**
    Method to get the hash of the tweet.
  - **Returns** – String with the hash of the tweet.

- **getId**

```
public java.lang.String getId()
```

  - **Description**
    Method to get the id of the tweet.
  - **Returns** – String with the id of the tweet.

- **getText**

```
public java.lang.String getText()
```

  - **Description**
    Method to get the text of the tweet.
  - **Returns** – String with the text of the tweet.

- **getUsername**

```
public java.lang.String getUsername()
```

  - **Description**
    Method to get the username of the tweet.
  - **Returns** – String with the username of the tweet.

- **toString**

```
public java.lang.String toString()
```

  - **Description**
    Method to get the string of the Tweet.
  - **Returns** – String with the String of the Tweet.

5.2.8. Class TweetCorrected

Tweet corrected class with the structure of a corrected tweet.

5.2.9. Declaration

```
public class TweetCorrected
 extends com.jmorenov.tweetsccore.twitter.Tweet
```

5.2.10. Constructor summary

TweetCorrected() Default constuctor of the class.
TweetCorrected(String) Constructor of the class.
TweetCorrected(String, String, String, String, String) Constructor of the class.
TweetCorrected(Tweet) Constructor from tweet.

5.2.11. Method summary

computeCorrectedText() Method to set the corrected text from the OOV words.
getCorrectedText() Method to get the corrected tweet.
getOOVWords() Method to get the Out-Of-Vocabulary words of the tweet.
setCorrectedText(String) Method to set the corrected tweet.
setOOVWords(List) Method to set the Out-Of-Vocabulary words of the tweet.

toString() Method to get the string of the Tweet Corrected.
toTweetNormString() Method to get the corrected text for Tweet Norm 2013.

5.2.12. Constructors

- **TweetCorrected**

```
public TweetCorrected()
```

  - **Description**
    Default constuctor of the class.

- **TweetCorrected**

```
public TweetCorrected(java.lang.String text)
```

  - **Description**
    Constructor of the class.
  - **Parameters**
    - `text` – of the tweet

- **TweetCorrected**

```
public TweetCorrected(java.lang.String id,java.lang.
    String username,java.lang.String hash,java.lang.
    String text,java.lang.String date)
```

  - **Description**
    Constructor of the class.
  - **Parameters**
    - `id` – of the tweet
    - `username` –
    - `hash` – of the tweet
    - `text` – of the tweet
    - `date` – of the tweet

- **TweetCorrected**

```
public TweetCorrected(Tweet tweet)
```

- **Description**
  Constructor from tweet.
- **Parameters**
  - `tweet` – Tweet

5.2.13. Methods

- **computeCorrectedText**

```
public void computeCorrectedText()
```

  - **Description**
    Method to set the corrected text from the OOV words.

- **getCorrectedText**

```
public java.lang.String getCorrectedText()
```

  - **Description**
    Method to get the corrected tweet.
  - **Returns** – String the corrected text

- **getOOVWords**

```
public java.util.List getOOVWords()
```

  - **Description**
    Method to get the Out-Of-Vocabulary words of the tweet.
  - **Returns** – List of OOV

- **setCorrectedText**

```
public void setCorrectedText(java.lang.String
    correctedText)
```

  - **Description**
    Method to set the corrected tweet.
  - **Parameters**
    - `correctedText` – the corrected text

- **setOOVWords**

  ```
  public void setOOVWords(java.util.List OOVWords)
  ```

  - **Description**
    Method to set the Out-Of-Vocabulary words of the tweet.
  - **Parameters**
    - `OOVWords` – the list of OOV

- **toString**

  ```
  public java.lang.String toString()
  ```

  - **Description**
    Method to get the string of the Tweet Corrected.
  - **Returns** – String with the String of the class

- **toTweetNormString**

  ```
  public java.lang.String toTweetNormString()
  ```

  - **Description**
    Method to get the corrected text for Tweet Norm 2013.
  - **Returns** – String with the correctec text.

5.2.14. Members inherited from class Tweet

`com.jmorenov.tweetsccore.twitter.Tweet` (in 5.2.1, page 26)

- `public String` **`getDate()`**
- `public String` **`getHash()`**
- `public String` **`getId()`**
- `public String` **`getText()`**
- `public String` **`getUsername()`**
- `public String` **`toString()`**

5.2.15. Class TwitterConfiguration

TwitterConfiguration class with the configuration of the connection with the API of twitter.

5.2.16. Declaration

```
public class TwitterConfiguration
 extends java.lang.Object
```

5.2.17. Method summary

5.2.18. Methods

- **getInstance**

  ```
  public static TwitterConfiguration getInstance()
  ```

  - **Description**
    Method to get the instance of the class.
  - **Returns** – TwitterConfiguration the instance of the class

- **getTwitterAccess**

  ```
  public Twitter getTwitterAccess()
  ```

  - **Description**
    Method to get the access to the API.
  - **Returns** – Twitter

5.3. Package com.jmorenov.tweetsccore.twitter.api

Package Contents *Page*

Classes

5.3.1. Class Search

Search class to do call on the Twitter API about tweets.

5.3.2. Declaration

```
public class Search
 extends java.lang.Object
```

5.3.3. Constructor summary

Search() Default constructor of the class.

5.3.4. Method summary

getAllTweetsOfUser(String) Method to get all the tweets of an user.
getTweetById(String) Method to search tweets by it id.
getTweetsByText(String) Method to search tweets.
getTweetsByTextOfUser(String, String) Method to search tweets of an user.

5.3.5. Constructors

- **Search**

  ```
  public Search()
  ```

  - **Description**
    Default constructor of the class.

5.3.6. Methods

- **getAllTweetsOfUser**

  ```
  public java.util.List getAllTweetsOfUser(java.lang.
      String username)
  ```

  - **Description**
    Method to get all the tweets of an user.
  - **Parameters**
    - `username` – the user
  - **Returns** – List of Status with the tweets

- **getTweetById**

  ```
  public com.jmorenov.tweetsccore.twitter.Tweet
      getTweetById(java.lang.String id)
  ```

- **Description**
  Method to search tweets by it id.
- **Parameters**
  - `id` – the id of the tweet
- **Returns** – Status The tweet

▪ **getTweetsByText**

```
public java.util.List getTweetsByText(java.lang.String
    text)
```

- **Description**
  Method to search tweets.
- **Parameters**
  - `text` – the text to search
- **Returns** – List of Status with the tweets

▪ **getTweetsByTextOfUser**

```
public java.util.List getTweetsByTextOfUser(java.lang.
    String username,java.lang.String text)
```

- **Description**
  Method to search tweets of an user.
- **Parameters**
  - `username` – the user
  - `text` – the text to search
- **Returns** – List of Status with the tweets

5.4. Package com.jmorenov.tweetsccore.ner

Package Contents *Page*

Classes

5.4.1. Class NER

POSTagging abstract class.

5.4.2. Declaration

```
public abstract class NER
 extends java.lang.Object
```

5.4.3. All known subclasses

StanfordNLPNER (in 5.4.14, page 38)

5.4.4. Constructor summary

NER()

5.4.5. Method summary

getNERElements(String) Method to get a list with the NER Elements detected.

5.4.6. Constructors

- **NER**

```
public NER()
```

5.4.7. Methods

- **getNERElements**

```
public abstract java.util.List getNERElements(java.lang.
    String text)
```

  - **Description**
    Method to get a list with the NER Elements detected.
  - **Returns** – List with the NER Elements.

5.4.8. Class NERELement

NERELement class.

5.4.9. Declaration

```
public class NERELement
 extends java.lang.Object
```

5.4.10. Constructor summary

NERELement(String, String) Constructor of the class

5.4.11. Method summary

getNerDetected() Method to get the ner detected.
getOriginalElement() Method to get the original element.

5.4.12. Constructors

- **NERELement**

  ```
  public NERELement(java.lang.String originalElement,java.
      lang.String nerDetected)
  ```

  - **Description**
    Constructor of the class
  - **Parameters**
    - `originalElement` –
    - `nerDetected` –

5.4.13. Methods

- **getNerDetected**

  ```
  public java.lang.String getNerDetected()
  ```

  - **Description**
    Method to get the ner detected.
  - **Returns** – String with the ner detected

- **getOriginalElement**

  ```
  public java.lang.String getOriginalElement()
  ```

  - **Description**
    Method to get the original element.
  - **Returns** – String with the original element

5.4.14. Class StanfordNLPNER

StanfordNLP class.

5.4.15. Declaration

```
public class StanfordNLPNER
 extends com.jmorenov.tweetsccore.ner.NER
```

5.4.16. Constructor summary

StanfordNLPNER(String)

5.4.17. Method summary

getNERElements(String) Method to get a list with the NER Elements detected.

5.4.18. Constructors

- **StanfordNLPNER**

```
public StanfordNLPNER(java.lang.String text)
```

5.4.19. Methods

- **getNERElements**

```
public java.util.List getNERElements(java.lang.String
    text)
```

  - **Description**
    Method to get a list with the NER Elements detected.
  - **Returns** – List with the NER Elements.

5.4.20. Members inherited from class NER

`com.jmorenov.tweetsccore.ner.NER` (in 5.4.1, page 36)

- `public abstract List` **`getNERElements`**`(java.lang.String` **`text`**`)`

5.5. Package com.jmorenov.tweetsccore.analyzer

Package Contents *Page*

Classes

5.5.1. Class AnalysisElement

AnalysisElement class.

5.5.2. Declaration

```
public class AnalysisElement
 extends java.lang.Object
```

5.5.3. Constructor summary

AnalysisElement(String, String, String, String, boolean) Constructor of the class.

5.5.4. Method summary

getForm() Method to get the form.
getLemma() Method to get the lemma.
getSenses() Method to get the senses.
getTag() Method to get the tag.
isMultiWord() Method to get if the element is multi word.
toString() Method to get the class as String.

5.5.5. Constructors

- **AnalysisElement**

```
public AnalysisElement(java.lang.String form,java.lang.
    String lemma,java.lang.String tag,java.lang.String
    senses,boolean isMultiWord)
```

  - **Description**
    Constructor of the class.

- **Parameters**
  - `form` –
  - `lemma` –
  - `tag` –
  - `isMultiWord` –

5.5.6. Methods

- **getForm**

  ```
  public java.lang.String getForm()
  ```

  - **Description**
    Method to get the form.
  - **Returns** – String with the form

- **getLemma**

  ```
  public java.lang.String getLemma()
  ```

  - **Description**
    Method to get the lemma.
  - **Returns** – String with the lemma

- **getSenses**

  ```
  public java.lang.String getSenses()
  ```

  - **Description**
    Method to get the senses.
  - **Returns** – String with the senses

- **getTag**

  ```
  public java.lang.String getTag()
  ```

  - **Description**
    Method to get the tag.
  - **Returns** – String with the tag

- **isMultiWord**

```
public boolean isMultiWord()
```

  - **Description**
    Method to get if the element is multi word.
  - **Returns** – Boolean with the test

- **toString**

```
public java.lang.String toString()
```

  - **Description**
    Method to get the class as String.
  - **Returns** – String

5.5.7. Class Analyzer

Analyzer abstract class.

5.5.8. Declaration

```
public abstract class Analyzer
 extends java.lang.Object
```

5.5.9. All known subclasses

FreelingAnalyzer (in 5.5.14, page 42)

5.5.10. Constructor summary

Analyzer()

5.5.11. Method summary

analyzeText(String) Method to get a list with the Analysis Elements.

5.5.12. Constructors

- **Analyzer**

```
public Analyzer()
```

5.5.13. Methods

- **analyzeText**

```
public abstract java.util.List analyzeText(java.lang.
    String text)
```

  - **Description**
    Method to get a list with the Analysis Elements.
  - **Returns** – List with the Analysis Elements.

5.5.14. Class FreelingAnalyzer

FreelingTokenizer class to tokenize a text.

5.5.15. Declaration

```
public class FreelingAnalyzer
 extends com.jmorenov.tweetsccore.analyzer.Analyzer
```

5.5.16. Constructor summary

FreelingAnalyzer() Constructor of the class.

5.5.17. Method summary

analyzeText(String) Method to get a list with the NER Elements detected.

5.5.18. Constructors

- **FreelingAnalyzer**

```
public FreelingAnalyzer()
```

  - **Description**
    Constructor of the class.

5.5.19. Methods

- **analyzeText**

```
public java.util.List analyzeText(java.lang.String text)
```

- **Description**
  Method to get a list with the NER Elements detected.
- **Returns** – List with the NER Elements.

5.5.20. Members inherited from class Analyzer

`com.jmorenov.tweetsccore.analyzer.Analyzer` (in 5.5.7, page 41)

- `public abstract List` **analyzeText**`(java.lang.String text)`

5.6. Package com.jmorenov.tweetsccore.candidates

Package Contents *Page*

Classes

5.6.1. Class Candidate

Candidate class that define the candidate element.

5.6.2. Declaration

```
public class Candidate
 extends java.lang.Object
```

5.6.3. Constructor summary

Candidate(String, String) Constructor of the class.

5.6.4. Method summary

getCandidate() Method to get the candidate.
getGeneratedBy() Method to get the method that generated the candidate.

5.6.5. Constructors

- **Candidate**

```
public Candidate(java.lang.String candidate,java.lang.
    String generatedBy)
```

  - **Description**
    Constructor of the class.
  - **Parameters**
    - `candidate` – String with the candidate
    - `generatedBy` – String with the method that generated the candidate

5.6.6. Methods

- **getCandidate**

```
public java.lang.String getCandidate()
```

  - **Description**
    Method to get the candidate.
  - **Returns** – String with the candidate

- **getGeneratedBy**

```
public java.lang.String getGeneratedBy()
```

  - **Description**
    Method to get the method that generated the candidate.
  - **Returns** – String with the generated by

5.6.7. Class CandidatesMethod

CandidatesMethod abstract class that define a method to generate candidates.

5.6.8. Declaration

```
public abstract class CandidatesMethod
 extends java.lang.Object
```

5.6.9. All known subclasses

MetaphoneCandidatesMethod (in 5.6.35, page 50), FastTextCandidatesMethod (in 5.6.21, page 47), LevenshteinFSTCandidatesMethod (in 5.6.28, page 48)

5.6.10. Constructor summary

CandidatesMethod()

5.6.11. Method summary

generateCandidates(OOV) Abstract method to generate condidates from an OOV.
getMethod() Abstract method to obtain the method description.

5.6.12. Constructors

- **CandidatesMethod**

```
public CandidatesMethod()
```

5.6.13. Methods

- **generateCandidates**

```
public abstract java.util.List generateCandidates(com.
    jmorenov.tweetsccore.extra.OOV oov)
```

  - **Description**
    Abstract method to generate condidates from an OOV.
  - **Parameters**
    - `oov` – OOV
  - **Returns** – List of Candidates

- **getMethod**

```
public abstract CandidatesMethodType getMethod()
```

  - **Description**
    Abstract method to obtain the method description.
  - **Returns** – CandidatesMethodType

5.6.14. Class CandidatesMethodType

CandidatesMethodType enum with the different methods to generate candidates.

5.6.15. Declaration

```
public final class CandidatesMethodType
 extends java.lang.Enum
```

5.6.16. Field summary

FastText
L_L
LevenshteinFST
Metaphone

5.6.17. Method summary

valueOf(String)
values()

5.6.18. Fields

- `public static final CandidatesMethodType` **LevenshteinFST**
- `public static final CandidatesMethodType` **Metaphone**
- `public static final CandidatesMethodType` **L_L**
- `public static final CandidatesMethodType` **FastText**

5.6.19. Methods

- **valueOf**

```
public static CandidatesMethodType valueOf(java.lang.
    String name)
```

- **values**

```
public static CandidatesMethodType[] values()
```

5.6.20. Members inherited from class Enum

`java.lang.Enum`

- `protected final Object` **clone**`() throws CloneNotSupportedException`
- `public final int` **compareTo**`(Enum` **arg0**`)`
- `public final boolean` **equals**`(Object` **arg0**`)`
- `protected final void` **finalize**`()`
- `public final Class` **getDeclaringClass**`()`
- `public final int` **hashCode**`()`
- `public final String` **name**`()`
- `public final int` **ordinal**`()`
- `public String` **toString**`()`
- `public static Enum` **valueOf**`(Class` **arg0**`, String` **arg1**`)`

5.6.21. Class FastTextCandidatesMethod

FastTextCandidatesMethod class that define a method to generate candidates.

5.6.22. Declaration

```
public class FastTextCandidatesMethod
 extends com.jmorenov.tweetsccore.candidates.
    CandidatesMethod
```

5.6.23. Constructor summary

FastTextCandidatesMethod() Constructor of the class.

5.6.24. Method summary

generateCandidates(OOV) Method to generate condidates from an OOV.
getMethod() Method to obtain the method description.

5.6.25. Constructors

- **FastTextCandidatesMethod**

```
public FastTextCandidatesMethod()
```

  - **Description**
    Constructor of the class.

5.6.26. Methods

- **generateCandidates**

```
public java.util.List generateCandidates(com.jmorenov.
    tweetsccore.extra.OOV oov)
```

  - **Description**
    Method to generate condidates from an OOV.
  - **Parameters**
    - `oov` – OOV
  - **Returns** – List of Candidates

- **getMethod**

```
public CandidatesMethodType getMethod()
```

  - **Description**
    Method to obtain the method description.
  - **Returns** – CandidatesMethodType

5.6.27. Members inherited from class CandidatesMethod

`com.jmorenov.tweetsccore.candidates.CandidatesMethod` (in 5.6.7, page 44)

- `public abstract List` **`generateCandidates`**`(com.jmorenov.tweetsccore.extra.OOV` **`oov`**`)`
- `public abstract CandidatesMethodType` **`getMethod`**`()`

5.6.28. Class LevenshteinFSTCandidatesMethod

LevenshteinFSTCandidatesMethod class that define a method to generate candidates.

5.6.29. Declaration

```
public class LevenshteinFSTCandidatesMethod
 extends com.jmorenov.tweetsccore.candidates.
    CandidatesMethod
```

5.6.30. Constructor summary

LevenshteinFSTCandidatesMethod() Constructor of the class.

5.6.31. Method summary

generateCandidates(OOV) Method to generate condidates from an OOV.
getMethod() Method to obtain the method description.

5.6.32. Constructors

- **LevenshteinFSTCandidatesMethod**

```
public LevenshteinFSTCandidatesMethod() throws java.lang.Exception
```

  - **Description**
    Constructor of the class.

5.6.33. Methods

- **generateCandidates**

```
public java.util.List generateCandidates(com.jmorenov.tweetsccore.extra.OOV oov)
```

  - **Description**
    Method to generate condidates from an OOV.
  - **Parameters**
    - `oov` – OOV
  - **Returns** – List of Candidates

- **getMethod**

```
public CandidatesMethodType getMethod()
```

  - **Description**
    Method to obtain the method description.
  - **Returns** – CandidatesMethodType

5.6.34. Members inherited from class CandidatesMethod

`com.jmorenov.tweetsccore.candidates.CandidatesMethod` (in 5.6.7, page 44)

- `public abstract List` **generateCandidates**`(com.jmorenov.tweetsccore.extra.OOV` **oov**`)`
- `public abstract CandidatesMethodType` **getMethod**`()`

5.6.35. Class MetaphoneCandidatesMethod

MetaphoneCandidatesMethod class that define a method to generate candidates.

5.6.36. Declaration

```
public class MetaphoneCandidatesMethod
 extends com.jmorenov.tweetsccore.candidates.
    CandidatesMethod
```

5.6.37. Constructor summary

MetaphoneCandidatesMethod() Constructor of the class.

5.6.38. Method summary

generateCandidates(OOV) Method to generate condidates from an OOV.
getMethod() Method to obtain the method description.

5.6.39. Constructors

- **MetaphoneCandidatesMethod**

```
public MetaphoneCandidatesMethod() throws java.io.
    IOException
```

  - **Description**
    Constructor of the class.

5.6.40. Methods

- **generateCandidates**

```
public java.util.List generateCandidates(com.jmorenov.
    tweetsccore.extra.OOV oov)
```

  - **Description**
    Method to generate condidates from an OOV.
  - **Parameters**
    - `oov` – OOV
  - **Returns** – List of Candidates

- **getMethod**

```
public CandidatesMethodType getMethod()
```

  - **Description**
    Method to obtain the method description.
  - **Returns** – CandidatesMethodType

5.6.41. Members inherited from class CandidatesMethod

`com.jmorenov.tweetsccore.candidates.CandidatesMethod` (in 5.6.7, page 44)

- `public abstract List` **generateCandidates**`(com.jmorenov.tweetsccore.extra.OOV oov)`
- `public abstract CandidatesMethodType` **getMethod**`()`

5.7. Package com.jmorenov.tweetsccore.evaluation

Package Contents *Page*

Classes

5.7.1. Class TweetNormEvaluationResult

5.7.2. Declaration

```
public class TweetNormEvaluationResult
 extends java.lang.Object
```

5.7.3. Constructor summary

TweetNormEvaluationResult(String)

5.7.4. Method summary

getAccurancy()
getErrors()
getNegatives()
getPositives()
getResultText()

5.7.5. Constructors

- **TweetNormEvaluationResult**

```
public TweetNormEvaluationResult(java.lang.String
    resultText)
```

5.7.6. Methods

- **getAccurancy**

```
public float getAccurancy()
```

- **getErrors**

```
public int getErrors()
```

- **getNegatives**

```
public int getNegatives()
```

- **getPositives**

```
public int getPositives()
```

- **getResultText**

```
public java.lang.String getResultText()
```

5.7.7. Class TweetNormEvaluator

TweetNormEvaluator class to evaluate methods of spell checker with Tweet Norm 2013 files to test.

5.7.8. Declaration

```
public class TweetNormEvaluator
 extends java.lang.Object
```

5.7.9. Constructor summary

TweetNormEvaluator() Default constructor of the class.
TweetNormEvaluator(String) Constructor of the class with parameter.
TweetNormEvaluator(String, boolean) Constructor of the class with parameters.

5.7.10. Method summary

evalutate(Method) Method to evaluate the defined file with a method of spell checker.
setAnnotatedFile(String) Method to define the file with the annotated tweets.
setEvaluatorScriptFile(String) Method to define the file of the evaluator script.
setIdsFile(String) Method to define the file with the ids of the tweets.
setResultFile(String) Method to define the result file.
setTweetsFile(String) Method to define the file with the tweets.
setWorkingDirectory(String) Method to define the working directory.

5.7.11. Constructors

- **TweetNormEvaluator**

```
public TweetNormEvaluator()
```

  - **Description**
    Default constructor of the class.

- **TweetNormEvaluator**

```
public TweetNormEvaluator(java.lang.String annotatedFile
    )
```

  - **Description**
    Constructor of the class with parameter.
  - **Parameters**
    - `annotatedFile` – String parameter with the name of the file with the annotated tweets.

- **TweetNormEvaluator**

```
public TweetNormEvaluator(java.lang.String annotatedFile
    ,boolean verbose)
```

- **Description**
  Constructor of the class with parameters.
- **Parameters**
  - `annotatedFile` – String parameter with the name of the file with the annotated tweets.
  - `verbose` – Boolean parameter to define the verbose control.

5.7.12. Methods

- **evalutate**

```
public TweetNormEvaluationResult evalutate(com.jmorenov.
    tweetsccore.method.Method method) throws java.io.
    IOException
```

- **Description**
  Method to evaluate the defined file with a method of spell checker.
- **Parameters**
  - `method` – parameter with the method to use.
- **Returns** – String with the output of the evaluation.
- **Throws**
  - `java.io.IOException` – when the file not found.
- **See also**
  - `com.jmorenov.tweetsccore.method.Method` (in 5.9.17, page 67)

- **setAnnotatedFile**

```
public void setAnnotatedFile(java.lang.String
    annotatedFile)
```

- **Description**
  Method to define the file with the annotated tweets.
- **Parameters**
  - `annotatedFile` – String parameter with the name of the file with the annotated tweets.

- **setEvaluatorScriptFile**

  ```
  public void setEvaluatorScriptFile(java.lang.String
      evaluatorScriptFile)
  ```

  - **Description**
    Method to define the file of the evaluator script.
  - **Parameters**
    - `evaluatorScriptFile` – String parameter with the name of the evaluator script.

- **setIdsFile**

  ```
  public void setIdsFile(java.lang.String idsFile)
  ```

  - **Description**
    Method to define the file with the ids of the tweets.
  - **Parameters**
    - `idsFile` – String parameter with the name of the file with the ids of the tweets.

- **setResultFile**

  ```
  public void setResultFile(java.lang.String resultFile)
  ```

  - **Description**
    Method to define the result file.
  - **Parameters**
    - `resultFile` – String parameter with the result file.

- **setTweetsFile**

  ```
  public void setTweetsFile(java.lang.String tweetsFile)
  ```

  - **Description**
    Method to define the file with the tweets.
  - **Parameters**
    - `tweetsFile` – String parameter with the name of the file with the tweets.

- **setWorkingDirectory**

```
public void setWorkingDirectory(java.lang.String
    workingDirectory)
```

  - **Description**

    Method to define the working directory.

  - **Parameters**

    - `workingDirectory` – String parameter with the working directory.

5.8. Package com.jmorenov.tweetsccore.extra

Package Contents *Page*

Classes

5.8.1. Class Annotation

Anotation enum with the different anotations posibilities of a tweet.

5.8.2. Declaration

```
public final class Annotation
 extends java.lang.Enum
```

5.8.3. Field summary

Correct
NoEs
value
Variation

5.8.4. Method summary

valueOf(String)
values()

5.8.5. Fields

- `public static final Annotation` **Variation**
- `public static final Annotation` **Correct**
- `public static final Annotation` **NoEs**
- `public int` **value**

5.8.6. Methods

- **valueOf**

```
public static Annotation valueOf(java.lang.String name)
```

- **values**

```
public static Annotation[] values()
```

5.8.7. Members inherited from class Enum

`java.lang.Enum`

- `protected final Object` **clone**`() throws CloneNotSupportedException`
- `public final int` **compareTo**`(Enum` **arg0**`)`
- `public final boolean` **equals**`(Object` **arg0**`)`
- `protected final void` **finalize**`()`
- `public final Class` **getDeclaringClass**`()`
- `public final int` **hashCode**`()`
- `public final String` **name**`()`
- `public final int` **ordinal**`()`
- `public String` **toString**`()`
- `public static Enum` **valueOf**`(Class` **arg0**`, String` **arg1**`)`

5.8.8. Class File

File class with funcionality to files.

5.8.9. Declaration

```
public class File
 extends java.lang.Object
```

5.8.10. Constructor summary

File()

5.8.11. Method summary

getStreamFromResources(String) Method to read a file stream from resources.
readToByte(String) Method to read a file to byte from resources.
readToStringArray(String) Method to read a file to array of string.

5.8.12. Constructors

- **File**

```
public File()
```

5.8.13. Methods

- **getStreamFromResources**

```
public static java.io.InputStream getStreamFromResources
    (java.lang.String fileName) throws java.io.
    IOException
```

  - **Description**
    Method to read a file stream from resources.
  - **Parameters**
    - `fileName` – the name of the file.
  - **Returns** – InputStream of the file.
  - **Throws**
    - `java.io.IOException` – when the file is not found.

- **readToByte**

```
public static byte[] readToByte(java.lang.String
    fileName) throws java.io.IOException
```

- **Description**
  Method to read a file to byte from resources.
- **Parameters**
  - `fileName` – the name of the file.
- **Returns** – byte[] of the file.
- **Throws**
  - `java.io.IOException` – when the file is not found.

- **readToStringArray**

```
public static java.lang.String[] readToStringArray(java.
    lang.String fileName) throws java.io.IOException
```

- **Description**
  Method to read a file to array of string.
- **Parameters**
  - `fileName` – String with the name of the file.
- **Returns** – String[] with the lines.
- **Throws**
  - `java.io.IOException` – when the file is not found.

5.8.14. Class FreelingInitializator

FreelingTokenizer class to initialize Freeling.

5.8.15. Declaration

```
public class FreelingInitializator
 extends java.lang.Object
```

5.8.16. Constructor summary

FreelingInitializator()

5.8.17. Method summary

init() Constructor of the class

5.8.18. Constructors

- **FreelingInitializator**

```
public FreelingInitializator()
```

5.8.19. Methods

- **init**

```
public static java.lang.String init()
```

  - **Description**
    Constructor of the class

5.8.20. Class OOV

Out-Of-Vocabulary class with the structure of a OOV word.

5.8.21. Declaration

```
public class OOV
 extends java.lang.Object
```

5.8.22. Constructor summary

OOV(String, int, int) Constructor of the class.

5.8.23. Method summary

getAnnotation()
getCorrection()
getEndPosition()
getStartPosition()
getToken()
setAnnotation(Annotation)
setCorrection(String)

5.8.24. Constructors

- **OOV**

```
public OOV(java.lang.String token, int startPosition, int
    endPosition)
```

  - **Description**
    Constructor of the class.
  - **Parameters**
    - `token` – String with the word or token of the OOV.

- `startPosition` – int with the initial position of the OOV in the original text.
- `endPosition` – int with the final position of the OOV in the original text.

5.8.25. Methods

- **getAnnotation**

```
public Annotation getAnnotation()
```

- **getCorrection**

```
public java.lang.String getCorrection()
```

- **getEndPosition**

```
public int getEndPosition()
```

- **getStartPosition**

```
public int getStartPosition()
```

- **getToken**

```
public java.lang.String getToken()
```

- **setAnnotation**

```
public void setAnnotation(Annotation annotation)
```

- **setCorrection**

```
public void setCorrection(java.lang.String correction)
```

5.8.26. Class Parser

5.8.27. Declaration

```
public class Parser
 extends java.lang.Object
```

5.8.28. Constructor summary

Parser()

5.8.29. Method summary

getHashtagRegex() Method to get the hashtag regex pattern.
getURLRegex() Method to get the url regex pattern.
getUsernameRegex() Method to get the user name regex pattern.
isHashtag(String) Method to check if a word is a hashtag of Twitter.
isPunctuationSign(String) Method to check if a word is a punctuation sign.
isUrl(String) Method to check if a word is a Url.
isUsername(String) Method to check if a word is a username of Twitter.
isValidWord(String) Method to check if a word is a valid word.
removeEmojiFromText(String) Method to remove the emojis from a text.

5.8.30. Constructors

- **Parser**

```
public Parser()
```

5.8.31. Methods

- **getHashtagRegex**

```
public static java.lang.String getHashtagRegex()
```

  - **Description**
    Method to get the hashtag regex pattern.
  - **Returns** – String with the pattern

- **getURLRegex**

```
public static java.lang.String getURLRegex()
```

  - **Description**
    Method to get the url regex pattern.
  - **Returns** – String with the pattern

- **getUsernameRegex**

```
public static java.lang.String getUsernameRegex()
```

- **Description**
  Method to get the user name regex pattern.
- **Returns** – String with the pattern

- **isHashtag**

```
public static java.lang.Boolean isHashtag(java.lang.
    String word)
```

  - **Description**
    Method to check if a word is a hashtag of Twitter.
  - **Parameters**
    - `word` – String with the word to check.
  - **Returns** – Boolean control parameter.

- **isPunctuationSign**

```
public static java.lang.Boolean isPunctuationSign(java.
    lang.String word)
```

  - **Description**
    Method to check if a word is a punctuation sign.
  - **Parameters**
    - `word` – String with the word to check.
  - **Returns** – Boolean control parameter.

- **isUrl**

```
public static java.lang.Boolean isUrl(java.lang.String
    word)
```

  - **Description**
    Method to check if a word is a Url.
  - **Parameters**
    - `word` – String with the word to check.
  - **Returns** – Boolean control parameter.

- **isUsername**

```
public static java.lang.Boolean isUsername(java.lang.
    String word)
```

- **Description**
  Method to check if a word is a username of Twitter.
- **Parameters**
  - `word` – String with the word to check.
- **Returns** – Boolean control parameter.

- **isValidWord**

```
public static java.lang.Boolean isValidWord(java.lang.
    String word)
```

- **Description**
  Method to check if a word is a valid word.
- **Parameters**
  - `word` – String with the word to check.
- **Returns** – Boolean control parameter.

- **removeEmojiFromText**

```
public static java.lang.String removeEmojiFromText(java.
    lang.String text)
```

- **Description**
  Method to remove the emojis from a text.
- **Parameters**
  - `text` – String with the text to remove the emojis.
- **Returns** – String with the text without the emojis.

5.9. Package com.jmorenov.tweetsccore.method

Package Contents *Page*

Classes

5.9.1. Class DictionaryAnalysisMethod

DictionaryAnalysisMethod class with the method of spell checker with dictionaries with analysis.

5.9.2. Declaration

```
public class DictionaryAnalysisMethod
 extends com.jmorenov.tweetsccore.method.DictionaryMethod
```

5.9.3. Constructor summary

DictionaryAnalysisMethod() Default constructor of the class.

5.9.4. Method summary

toString() Method to get the String of the method.

5.9.5. Constructors

- **DictionaryAnalysisMethod**

```
public DictionaryAnalysisMethod() throws java.io.
    IOException
```

  - **Description**
    Default constructor of the class.
  - **Throws**
    - `java.io.IOException` – when the file not found.

5.9.6. Methods

- **toString**

```
public java.lang.String toString()
```

  - **Description**
    Method to get the String of the method.
  - **Returns** – String with the String of the method.

5.9.7. Members inherited from class DictionaryMethod

`com.jmorenov.tweetsccore.method.DictionaryMethod` (in 5.9.9, page 66)

- `public TweetCorrected` **correctTweet**`(com.jmorenov.tweetsccore.twitter.Tweet` **tweet**`)`
- `public String` **toString**`()`

5.9.8. Members inherited from class Method

`com.jmorenov.tweetsccore.method.Method` (in 5.9.17, page 67)

- `public abstract TweetCorrected` **correctTweet**`(com.jmorenov.tweetsccore.twitter.Tweet` **tweet**`)`
- `public abstract String` **toString**`()`

5.9.9. Class DictionaryMethod

DictionaryMethod class with the method of spell checker with dictionaries.

5.9.10. Declaration

```
public class DictionaryMethod
 extends com.jmorenov.tweetsccore.method.Method
```

5.9.11. All known subclasses

DictionaryAnalysisMethod (in 5.9.1, page 65)

5.9.12. Constructor summary

DictionaryMethod() Default constructor of the class.

5.9.13. Method summary

correctTweet(Tweet) Method to get the corrected tweet.
toString() Method to get the String of the method.

5.9.14. Constructors

- **DictionaryMethod**

```
public DictionaryMethod() throws java.io.IOException
```

  - **Description**
    Default constructor of the class.
  - **Throws**
    - `java.io.IOException` – when the file not found.

5.9.15. Methods

- **correctTweet**

```
public com.jmorenov.tweetsccore.twitter.TweetCorrected
    correctTweet(com.jmorenov.tweetsccore.twitter.Tweet
    tweet)
```

  - **Description**
    Method to get the corrected tweet.
  - **Parameters**
    - `tweet` – Tweet with the tweet to correct.
  - **Returns** – TweetCorrected with the corrected tweet.

- **toString**

```
public java.lang.String toString()
```

  - **Description**
    Method to get the String of the method.
  - **Returns** – String with the String of the method.

5.9.16. Members inherited from class Method

`com.jmorenov.tweetsccore.method.Method` (in 5.9.17, page 67)

- `public abstract TweetCorrected` **correctTweet**`(com.jmorenov.tweetsccore.twitter.Tw`
  tweet)
- `public abstract String` **toString**`()`

5.9.17. Class Method

Method abstract class.

5.9.18. Declaration

```
public abstract class Method
 extends java.lang.Object
```

5.9.19. All known subclasses

DictionaryAnalysisMethod (in 5.9.1, page 65), DictionaryMethod (in 5.9.9, page 66)

5.9.20. Constructor summary

Method() Default constructor of the class.

5.9.21. Method summary

correctTweet(Tweet) Abstract method to get the corrected tweet.
toString() Abstract method to get the String of the method.

5.9.22. Constructors

- **Method**

```
public Method()
```

  - **Description**
    Default constructor of the class.

5.9.23. Methods

- **correctTweet**

```
public abstract com.jmorenov.tweetsccore.twitter.
    TweetCorrected correctTweet(com.jmorenov.tweetsccore.
    twitter.Tweet tweet)
```

  - **Description**
    Abstract method to get the corrected tweet.
  - **Parameters**
    - `tweet` – Tweet with the tweet to correct.
  - **Returns** – TweetCorrected with the corrected tweet.

- **toString**

```
public abstract java.lang.String toString()
```

  - **Description**
    Abstract method to get the String of the method.
  - **Returns** – String with the String of the method.

5.10. Package com.jmorenov.tweetsccore.post

Package Contents *Page*

Classes

5.10.1. Class FreeLingPOST

FreeLingPOST class to get the POST of a text. https://talp-upc.gitbooks.io/freeling-4-1-user-manual/content/

5.10.2. Declaration

```
public class FreeLingPOST
 extends java.lang.Object
```

5.10.3. Constructor summary

FreeLingPOST()

5.10.4. Constructors

- **FreeLingPOST**

```
public FreeLingPOST()
```

5.10.5. Class OpenNLPPOST

OpenNLPPOST class to get the POST of a text. https://opennlp.apache.org/docs/1.8.4/ma

5.10.6. Declaration

```
public class OpenNLPPOST
 extends java.lang.Object
```

5.10.7. Constructor summary

OpenNLPPOST()

5.10.8. Constructors

- **OpenNLPPOST**

```
public OpenNLPPOST()
```

5.10.9. Class POST

5.10.10. Declaration

```
public abstract class POST
 extends java.lang.Object
```

5.10.11. All known subclasses

StanfordNLPPOST (in 5.10.16, page 70)

5.10.12. Constructor summary

POST()

5.10.13. Method summary

getTags()

5.10.14. Constructors

- **POST**

```
public POST()
```

5.10.15. Methods

- **getTags**

```
public abstract java.lang.String getTags()
```

5.10.16. Class StanfordNLPPOST

5.10.17. Declaration

```
public class StanfordNLPPOST
 extends com.jmorenov.tweetsccore.post.POST
```

5.10.18. Constructor summary

StanfordNLPPOST(String)

5.10.19. Method summary

getTags()

5.10.20. Constructors

- **StanfordNLPPOST**

```
public StanfordNLPPOST(java.lang.String text)
```

5.10.21. Methods

- **getTags**

```
public abstract java.lang.String getTags()
```

5.10.22. Members inherited from class POST

`com.jmorenov.tweetsccore.post.POST` (in 5.10.9, page 70)

- `public abstract String getTags()`

5.11. Package com.jmorenov.tweetsccore.spellchecker

Package Contents *Page*

Classes

SpellChecker 71
SpellChecker class to correct a text.

5.11.1. Class SpellChecker

SpellChecker class to correct a text.

5.11.2. Declaration

```
public class SpellChecker
 extends java.lang.Object
```

5.11.3. Constructor summary

SpellChecker(Method) Constructor of the class.

5.11.4. Method summary

correctText(String) Method to correct the text.
correctTweet(Tweet) Method to correct a tweet.
getMethodDescription() Method to get spell checker method description.
setMethod(Method) Method to define the spell checker method.

5.11.5. Constructors

- **SpellChecker**

```
public SpellChecker(com.jmorenov.tweetsccore.method.
    Method method)
```

  - **Description**
    Constructor of the class.
  - **Parameters**
    - `method` – parameter with the method to use.
  - **See also**
    - `com.jmorenov.tweetsccore.method.Method` (in 5.9.17, page 67)

5.11.6. Methods

- **correctText**

```
public java.lang.String correctText(java.lang.String
    text)
```

  - **Description**
    Method to correct the text.
  - **Parameters**
    - `text` – String with the text to correct.
  - **Returns** – String with the corrected text.

- **correctTweet**

```
public com.jmorenov.tweetsccore.twitter.TweetCorrected
    correctTweet(com.jmorenov.tweetsccore.twitter.Tweet
    tweet)
```

- **Description**
  Method to correct a tweet.
- **Parameters**
  - `tweet` – Tweet with the tweet.
- **Returns** – TweetCorrected with the corrected tweet.

▪ **getMethodDescription**

```
public java.lang.String getMethodDescription()
```

- **Description**
  Method to get spell checker method description.
- **Returns** – String with the description of the method.

▪ **setMethod**

```
public void setMethod(com.jmorenov.tweetsccore.method.
    Method method)
```

- **Description**
  Method to define the spell checker method.
- **Parameters**
  - `method` – parameter with the method to use.

5.12. Package com.jmorenov.tweetsccore.tokenizer

Package Contents *Page*

Classes

5.12.1. Class FreelingTokenizer

FreelingTokenizer class to tokenize a text.

5.12.2. Declaration

```
public class FreelingTokenizer
 extends com.jmorenov.tweetsccore.tokenizer.Tokenizer
```

5.12.3. Constructor summary

FreelingTokenizer() Constructor of the class

5.12.4. Method summary

getTokens(String) Method to get the tokens from the text.

5.12.5. Constructors

- **FreelingTokenizer**

  ```
  public FreelingTokenizer()
  ```

  - **Description**
    Constructor of the class

5.12.6. Methods

- **getTokens**

  ```
  public java.util.List getTokens(java.lang.String text)
  ```

  - **Description**
    Method to get the tokens from the text.
  - **Parameters**
    - `text` – String with the text
  - **Returns** – List of String with the tokens

5.12.7. Members inherited from class Tokenizer

`com.jmorenov.tweetsccore.tokenizer.Tokenizer` (in 5.12.29, page 78)

- `public abstract List` **`getTokens`**`(java.lang.String text)`

5.12.8. Class NGramTokenizer

NGramTokenizer class to tokenize a text. https://opennlp.apache.org/docs/1.8.4/manual/op

5.12.9. Declaration

```
public class NGramTokenizer
 extends com.jmorenov.tweetsccore.tokenizer.Tokenizer
```

5.12.10. Constructor summary

NGramTokenizer() Constructor of the class

5.12.11. Method summary

getTokens(String) Method to get the tokens from the text.

5.12.12. Constructors

- **NGramTokenizer**

  ```
  public NGramTokenizer()
  ```

  - **Description**
    Constructor of the class

5.12.13. Methods

- **getTokens**

  ```
  public java.util.List getTokens(java.lang.String text)
  ```

  - **Description**
    Method to get the tokens from the text.
  - **Parameters**
    - `text` – String with the text
  - **Returns** – List of String with the tokens

5.12.14. Members inherited from class Tokenizer

`com.jmorenov.tweetsccore.tokenizer.Tokenizer` (in 5.12.29, page 78)

- `public abstract List` **getTokens**`(java.lang.String` **text**`)`

5.12.15. Class OpenNLPTokenizer

OpenNLPTokenizer class to tokenize a text. https://opennlp.apache.org/docs/1.8.4/manual/opennlp.ht

5.12.16. Declaration

```
public class OpenNLPTokenizer
 extends com.jmorenov.tweetsccore.tokenizer.Tokenizer
```

5.12.17. Constructor summary

OpenNLPTokenizer() Constructor of the class

5.12.18. Method summary

getTokens(String) Method to get the tokens from the text.

5.12.19. Constructors

- **OpenNLPTokenizer**

```
public OpenNLPTokenizer() throws java.io.IOException
```

  - **Description**
    Constructor of the class
  - **Throws**
    - `java.io.IOException` –

5.12.20. Methods

- **getTokens**

```
public java.util.List getTokens(java.lang.String text)
```

  - **Description**
    Method to get the tokens from the text.
  - **Parameters**
    - `text` – String with the text
  - **Returns** – List of String with the tokens

5.12.21. Members inherited from class Tokenizer

`com.jmorenov.tweetsccore.tokenizer.Tokenizer` (in 5.12.29, page 78)

- `public abstract List` **`getTokens`**`(java.lang.String` **`text`**`)`

5.12.22. Class StanfordNLPTokenizer

StanfordNLPTokenizer class to tokenize a text. https://stanfordnlp.github.io/CoreNLP/

5.12.23. Declaration

```
public class StanfordNLPTokenizer
 extends com.jmorenov.tweetsccore.tokenizer.Tokenizer
```

5.12.24. Constructor summary

StanfordNLPTokenizer() Constructor of the class

5.12.25. Method summary

getTokens(String) Method to get the tokens from the text.

5.12.26. Constructors

- **StanfordNLPTokenizer**

```
public StanfordNLPTokenizer() throws java.io.IOException
```

  - **Description**
    Constructor of the class

5.12.27. Methods

- **getTokens**

```
public java.util.List getTokens(java.lang.String text)
```

  - **Description**
    Method to get the tokens from the text.
  - **Parameters**
    - `text` – String with the text
  - **Returns** – List of String with the tokens

5.12.28. Members inherited from class Tokenizer

`com.jmorenov.tweetsccore.tokenizer.Tokenizer` (in 5.12.29, page 78)

- `public abstract List` **`getTokens`**`(java.lang.String` **`text`**`)`

5.12.29. Class Tokenizer

Tokenizer abstract class.

5.12.30. Declaration

```
public abstract class Tokenizer
 extends java.lang.Object
```

5.12.31. All known subclasses

FreelingTokenizer (in 5.12.1, page 74), StanfordNLPTokenizer (in 5.12.22, page 77), NGramTokenizer (in 5.12.8, page 75), OpenNLPTokenizer (in 5.12.15, page 76)

5.12.32. Constructor summary

Tokenizer()

5.12.33. Method summary

getTokens(String) Method to get the tokens from a text.

5.12.34. Constructors

- **Tokenizer**

```
public Tokenizer()
```

5.12.35. Methods

- **getTokens**

```
public abstract java.util.List getTokens(java.lang.
    String text)
```

  - **Description**
    Method to get the tokens from a text.
  - **Parameters**
    - `text` – String with the text
  - **Returns** – List of String with the tokens

6. TweetSCWeb Documentación del código (Java Documentation)

Class Hierarchy

Classes

- java.lang.Object
  - SpringBootServletInitializer
    - com.jmorenov.tweetscweb.ServletInitializer (in 6.1.13, page 82)
  - com.jmorenov.tweetscweb.Application (in 6.1.1, page 80)
  - com.jmorenov.tweetscweb.Response (in 6.1.7, page 80)
  - com.jmorenov.tweetscweb.TweetCorrectedListModel (in 6.1.19, page 83)
  - com.jmorenov.tweetscweb.TweetCorrectorAPIController (in 6.1.32, page 84)
  - com.jmorenov.tweetscweb.TweetCorrectorController (in 6.1.38, page 86)
  - com.jmorenov.tweetscweb.TweetListModel (in 6.1.44, page 87)
  - com.jmorenov.tweetscweb.TweetModel (in 6.1.50, page 88)
    - com.jmorenov.tweetscweb.TweetCorrectedModel (in 6.1.25, page 84)
  - com.jmorenov.tweetscweb.TweetSearchQuery (in 6.1.59, page 89)
  - com.jmorenov.tweetscweb.TweetSearchQueryModel (in 6.1.65, page 90)

6.1. Package com.jmorenov.tweetscweb

Package Contents *Page*

Classes

TweetSearchQuery class with the functionality to work over the queries.

TweetSearchQueryModel 90
TweetSearchQueryModel class with the model of the queries.

6.1.1. Class Application

Application class.

6.1.2. Declaration

```
public class Application
 extends java.lang.Object
```

6.1.3. Constructor summary

Application()

6.1.4. Method summary

main(String[]) Main method of the web application.

6.1.5. Constructors

- **Application**

```
public Application()
```

6.1.6. Methods

- **main**

```
public static void main(java.lang.String[] args)
```

  - **Description**
    Main method of the web application.
  - **Parameters**
    - `args` – String[] with the arguments of the execution.

6.1.7. Class Response

Response class.

6.1.8. Declaration

```
public class Response
 extends java.lang.Object
```

6.1.9. Constructor summary

Response() Default constructor of the class.
Response(String, Object) Constructor of the class.

6.1.10. Method summary

getData() Method to get the data of the response
getStatus() Method to get the status of the response.
setData(Object) Method to define the data of the response
setStatus(String) Method to define the status of the response.

6.1.11. Constructors

- **Response**

```
public Response()
```

  - **Description**
    Default constructor of the class.

- **Response**

```
public Response(java.lang.String status,java.lang.Object
    data)
```

  - **Description**
    Constructor of the class.
  - **Parameters**
    - `status` – String with the status of the response.
    - `data` – Object with the value of the response.

6.1.12. Methods

- **getData**

```
public java.lang.Object getData()
```

- **Description**
  Method to get the data of the response
- **Returns** – Object with the value of the response.

- **getStatus**

```
public java.lang.String getStatus()
```

  - **Description**
    Method to get the status of the response.
  - **Returns** – String with the status of the response.

- **setData**

```
public void setData(java.lang.Object data)
```

  - **Description**
    Method to define the data of the response
  - **Parameters**
    - `data` – Object with the value of the response.

- **setStatus**

```
public void setStatus(java.lang.String status)
```

  - **Description**
    Method to define the status of the response.
  - **Parameters**
    - `status` – String with the status.

6.1.13. Class ServletInitializer

ServletInitializer class.

6.1.14. Declaration

```
public class ServletInitializer
 extends SpringBootServletInitializer
```

6.1.15. Constructor summary

ServletInitializer()

6.1.16. Method summary

configure(SpringApplicationBuilder)

6.1.17. Constructors

- **ServletInitializer**

```
public ServletInitializer()
```

6.1.18. Methods

- **configure**

```
protected SpringApplicationBuilder configure(
    SpringApplicationBuilder application)
```

6.1.19. Class TweetCorrectedListModel

TweetCorrectedListModel class with the model of the list of tweets corrected.

6.1.20. Declaration

```
public class TweetCorrectedListModel
 extends java.lang.Object
```

6.1.21. Field summary

tweets

6.1.22. Constructor summary

TweetCorrectedListModel()

6.1.23. Fields

- `public java.util.List` **tweets**

6.1.24. Constructors

- **TweetCorrectedListModel**

```
public TweetCorrectedListModel()
```

6.1.25. Class TweetCorrectedModel

6.1.26. Declaration

```
public class TweetCorrectedModel
 extends com.jmorenov.tweetscweb.TweetModel
```

6.1.27. Field summary

correctedText

6.1.28. Constructor summary

TweetCorrectedModel()
TweetCorrectedModel(TweetCorrected)

6.1.29. Fields

- `public java.lang.String` **correctedText**

6.1.30. Constructors

- **TweetCorrectedModel**

  ```
  public TweetCorrectedModel()
  ```

- **TweetCorrectedModel**

  ```
  public TweetCorrectedModel(TweetCorrected tweetCorrected
      )
  ```

6.1.31. Members inherited from class TweetModel

`com.jmorenov.tweetscweb.TweetModel` (in 6.1.50, page 88)

- `public` **date**
- `public` **hash**
- `public` **id**
- `public` **text**
- `public Tweet` **toTweet()**
- `public` **username**

6.1.32. Class TweetCorrectorAPIController

TweetCorrectorAPIController class with the controller of the API.

6.1.33. Declaration

```
public class TweetCorrectorAPIController
 extends java.lang.Object
```

6.1.34. Constructor summary

TweetCorrectorAPIController()

6.1.35. Method summary

advancedCorrectSubmit(TweetListModel) Method to control the api calls of advanced corrector.

getTweets(TweetSearchQueryModel) Method to control the api calls.

simpleCorrectSubmit(TweetModel) Method to control the api calls of simple corrector.

6.1.36. Constructors

- **TweetCorrectorAPIController**

  ```
  public TweetCorrectorAPIController()
  ```

6.1.37. Methods

- **advancedCorrectSubmit**

  ```
  public Response advancedCorrectSubmit(TweetListModel
    tweetListModel)
  ```

  - **Description**
    Method to control the api calls of advanced corrector.
  - **Parameters**
    - `tweetListModel` – with the model of the call.
  - **Returns** – with the response of the call.

- **getTweets**

  ```
  public Response getTweets(TweetSearchQueryModel
    tweetSearchQueryModel)
  ```

  - **Description**
    Method to control the api calls.

- **Parameters**
  - `tweetSearchQueryModel` – with the model of the call.
- **Returns** – with the response of the call.

- **simpleCorrectSubmit**

```
public Response simpleCorrectSubmit(TweetModel
    tweetModel)
```

- **Description**
  Method to control the api calls of simple corrector.
- **Parameters**
  - `tweetModel` – with the model of the call.
- **Returns** – with the response of the call.

6.1.38. Class TweetCorrectorController

TweetCorrectorController class with the controller of the frontend.

6.1.39. Declaration

```
public class TweetCorrectorController
 extends java.lang.Object
```

6.1.40. Constructor summary

TweetCorrectorController()

6.1.41. Method summary

homeForm(Model) Method to control the frontend calls.

6.1.42. Constructors

- **TweetCorrectorController**

```
public TweetCorrectorController()
```

6.1.43. Methods

- **homeForm**

```
public java.lang.String homeForm(Model model)
```

  - **Description**
    Method to control the frontend calls.
  - **Parameters**
    - `model` – Model with the model of the call.
  - **Returns** – String with the template to show.

6.1.44. Class TweetListModel

TweetListModel class with the model of the list of tweets.

6.1.45. Declaration

```
public class TweetListModel
 extends java.lang.Object
```

6.1.46. Field summary

tweets

6.1.47. Constructor summary

TweetListModel()

6.1.48. Fields

- `public java.util.List` **tweets**

6.1.49. Constructors

- **TweetListModel**

```
public TweetListModel()
```

6.1.50. Class TweetModel

6.1.51. Declaration

```
public class TweetModel
 extends java.lang.Object
```

6.1.52. All known subclasses

TweetCorrectedModel (in 6.1.25, page 84)

6.1.53. Field summary

date
hash
id
text
username

6.1.54. Constructor summary

TweetModel()
TweetModel(Tweet)

6.1.55. Method summary

toTweet()

6.1.56. Fields

- `public java.lang.String` **id**
- `public java.lang.String` **username**
- `public java.lang.String` **hash**
- `public java.lang.String` **text**
- `public java.lang.String` **date**

6.1.57. Constructors

- **TweetModel**

  ```
  public TweetModel()
  ```

- **TweetModel**

```
public TweetModel(Tweet tweet)
```

6.1.58. Methods

- **toTweet**

```
public Tweet toTweet()
```

6.1.59. Class TweetSearchQuery

TweetSearchQuery class with the functionality to work over the queries.

6.1.60. Declaration

```
public class TweetSearchQuery
 extends java.lang.Object
```

6.1.61. Constructor summary

TweetSearchQuery(TweetSearchQueryModel) Constructor of the class.

6.1.62. Method summary

isValidQuery() Method to get if the query is valid or not.
loadTweets() Method to load the tweets from the query.

6.1.63. Constructors

- **TweetSearchQuery**

```
public TweetSearchQuery(TweetSearchQueryModel
    tweetSearchQueryModel)
```

  - **Description**

    Constructor of the class.

  - **Parameters**

    - `tweetSearchQueryModel` – TweetSearchQueryModel with the data.

6.1.64. Methods

- **isValidQuery**

```
public boolean isValidQuery()
```

  - **Description**
    Method to get if the query is valid or not.
  - **Returns** – Boolean

- **loadTweets**

```
public java.util.List loadTweets()
```

  - **Description**
    Method to load the tweets from the query.
  - **Returns** – List of tweet

6.1.65. Class TweetSearchQueryModel

TweetSearchQueryModel class with the model of the queries.

6.1.66. Declaration

```
public class TweetSearchQueryModel
 extends java.lang.Object
```

6.1.67. Constructor summary

TweetSearchQueryModel()

6.1.68. Method summary

getQuery() Method to get the query.
getTweets() Method to get the tweets.
setQuery(String) Method to set the query.
setTweets(List) Method to set the tweets.

6.1.69. Constructors

- **TweetSearchQueryModel**

```
public TweetSearchQueryModel()
```

6.1.70. Methods

- **getQuery**

  ```
  public java.lang.String getQuery()
  ```

  - **Description**
    Method to get the query.
  - **Returns** – String

- **getTweets**

  ```
  public java.util.List getTweets()
  ```

  - **Description**
    Method to get the tweets.
  - **Returns** – List of tweet

- **setQuery**

  ```
  public void setQuery(java.lang.String query)
  ```

  - **Description**
    Method to set the query.
  - **Parameters**
    - `query` – String

- **setTweets**

  ```
  public void setTweets(java.util.List tweets)
  ```

  - **Description**
    Method to set the tweets.
  - **Parameters**
    - `tweets` – List of tweet

7. TweetSCExecutable Documentación del código (Java Documentation)

Class Hierarchy

Classes

- java.lang.Object
  - com.jmorenov.tweetscexecutable.SpellCheckerRun (in 7.1.1, page 93)

7.1. Package com.jmorenov.tweetscexecutable

Package Contents *Page*

Classes

7.1.1. Class SpellCheckerRun

SpellCheckerRun class.

7.1.2. Declaration

```
public class SpellCheckerRun
 extends java.lang.Object
```

7.1.3. Constructor summary

SpellCheckerRun()

7.1.4. Method summary

main(String[]) Main method

7.1.5. Constructors

- **SpellCheckerRun**

```
public SpellCheckerRun()
```

7.1.6. Methods

- **main**

public static void main(java.lang.String[] args) **throws** java.io.IOException

- **Description**
  Main method
- **Parameters**
  - `args` – String[] with the arguments.
- **Throws**
  - `java.io.IOException` – when the files are not found.

8. Recursos utilizados

Los recursos utilizados por nuestro sistema son variados, desde diccionarios hasta bibliotecas para el desarrollo. Empezando desde el paquete TweetSCCore, se ha utilizado la biblioteca StanfordNLP [29] para la tokenización, además está disponible en el paquete la biblioteca Freeling [22]. El acceso a la API de twitter se realiza mediante la bibliteca para Java Twitter4j. La detección de OOV utiliza tres diccionarios, el diccionario de español proporcionado por la herramienta Aspell, el diccionario de entidades JRC y un diccionario de inglés [14]. Los métodos de la generación de candidatos utilizan los recursos: Algoritmo del metáfono [48], fastText [17] y la biblioteca liblevenshtein [15] para el FST. El ranking de candidatos utiliza la biblioteca [4] para el modelo del lenguaje N-Grama y la distancia Damerau-Levenshtein ofrecida por String.Util de java.

Además el corpus de tweets para la evaluación es el que recopiló Tweet-Norm 2013 [3] y su script en Python para los resultados.

El paquete TweetSCWeb utiliza el framework Spring Boot para la aplicación web junto con el sdk de Google Cloud para ofrecer la aplicación en la nube.

9. Evaluación

Esta sección describe la evaluación que se ha hecho de la solución propuesta. Primero se define la metodología utilizada para evaluar la solución, en segundo lugar explicamos el corpus utilizado como datos de entrada, posteriormente el gold standard actual y por último los experimentos que hemos realizada con sus resultados.

9.1. Metodología

La metodología que hemos seguido ha sido la misma que en la tarea compartida Tweet-Norm 2013 [69]. Ellos utilizan como medida de evaluación la corrección de errores, sólo tiene en cuenta si la forma propuesta es correcta en base a los criterios: **correcta** si la forma original era correcta y no se ha realizado ninguna normalización o si la forma original era incorrecta y el candidato seleccionado es el correcto; **errónea** en cualquier otro caso. La evaluación final es el número de decisiones realizadas correctamente sobre el total de palabras OOV.

9.2. Corpus

El corpus utilizado es el mismo que en la tarea compartida Tweet-Norm 2013 [69], en donde utilizan dos subconjuntos uno de desarrollo con 500 tweets y otro de evaluación con 600 tweets.

9.2.1. Gold Standard

Nuestro gold standard ha sido el sistema propuesto RAE [23] en Tweet-Norm 2013 [69] donde consiguieron un resultado de 0.781 de precisión. Su sistema se basa en trasductores de estados finitos con pesos.

9.3. Experimentos

(Experimentos realizados)

10. Conclusiones

El proyecto realizado está compuesto de una parte de investigación, cómo se demuestra con el estado del arte y las diferentes soluciones que se han ido realizando hasta llegar a nuestra solución final. Además de la otra parte de desarrollo e implementación de software, ofreciéndolo para todos en código abierto y en una aplicación web [46].

Este desarrollo software se ha dividido en tres componentes o módulos:

- TweetSCCore: Núcleo del proyecto con la funcionalidad para corregir textos de twitter.
- TweetSCWeb: Aplicación web para corregir textos.
- TweetSCExecutable: Ejecutable java para corregir textos desde línea de comandos.

Se puede concluir que nuestro objetivo era construir un corrector de texto para twitter (sección 1.2) y se ha conseguido cómo se ha demostrado en las secciones sección 3 y sección 4.

11. Líneas Futuras

Las líneas futuras son muy ámplias ya que estamos en un tema bastante reciente, sobre todo en español cómo se puede ver en el sección 2, y los resultados se pueden mejorar de bastantes formas. Centrándonos en nuestro sistema una mejora futura sería el añadir contexto a los tweets a partir de sus hashtag, usuarios, imágenes o emoticonos; de forma que se pudiera reducir el conjunto de candidatos o añadir nuevos a partir de estos datos. También se podría realizar un análisis de sentimientos sobre el tweet después de normalizar por si se pudiera mejorar la corrección de algún OOV.

ANEXOS

Glosario de términos

- **Modelo (estadístico) del lenguaje:** Un modelo estadístico del lenguaje es una distribución de probabilidad sobre secuencias de palabras. Un tipo de modelo del lenguaje es el unigrama, también se suele llamar modelo de bolsa de palabras. La dispersidad en los datos es un problema al construir modelos del lenguaje. La secuencia de palabras más probable puede no aparecer en los datos de entrenamiento. Una solución es realizar la suposición de que la probabilidad de una palabra sólo depende de las n palabras previas. Esto es conocido como el modelo n-grama, unigrama cuando n=1. Los modelos del lenguaje neuronales o modelos del lenguaje continuos: modelo del lenguaje Skip-gram, base de word2vec.

- **Modelo del lenguaje N-grama:** Un modelo de n-grama es un tipo de modelo probabilístico que permite hacer predicción estadística del próximo elemento de cierta secuencia de elementos sucedida hasta el momento. Un modelo de n-grama puede ser definido por una cadena de Márkov de orden n-1. Predice x_i basándose en los n elementos anteriores.

- **Cadena de Márkov:** En la teoría de la probabilidad, se conoce como cadena de Márkov o modelo de Márkov a un tipo especial de proceso estocástico discreto en el que la probabilidad de que ocurra un evento depende solamente del evento inmediatamente anterior. En matemáticas se define como un proceso estocástico discreto que cumple con la propiedad de Márkov, es decir, si se conoce la historia del sistema hasta su instante actual, su estado presente resume toda la información relevante para describir en probabilidad su futuro.

- **Proceso de Márkov:** Fenómeno aleatorio dependiente del tiempo para el cual se cumple la propiedad de Márkov. Frecuentemente el término cadena de Márkov se usa para dar a entender que un proceso de Márkov tiene un espacio de estados discreto (infinito o numerable).

- **Modelo oculto de Márkov:** Un modelo oculto de Márkov (Hidden Markov Model, HMM) es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Márkov de parámetros desconocidos. El objetivo es determinar los parámetros desconocidos (u ocultos) de dicha cadena a partir de los parámetros observables. Un HMM se puede considerar como la red bayesiana más simple.

- **Etiquetado gramatical:** El etiquetado gramatical (part-of-speech tagging, POS tagging o POST) se considera el proceso de asignar a cada palabra de un texto su categoría gramatical. Las soluciones se pueden dividir en dos grandes grupos: aproximaciones lingüísticas basadas en un conjunto de reglas establecidas manualmente por expertos aprendidas de forma (semi)automática, y

las aproximaciones de aprendizaje automático que usan textos, generalmente anotados, para establecer los modelos. Además se pueden encontrar aproximaciones híbridas que combinan ciertos aspectos de las anteriores.

Referencias

[1] Alicia Ageno, Pere R. Comas, Lluís Padró, and Jordi Turmo. The talp-upc approach to tweet-norm 2013, 2013.

[2] Iñaki Alegria, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. Tweetnorm: a benchmark for lexical normalization of spanish tweets, 2015.

[3] Iñaki Alegria, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. Introducción a la tarea compartida tweet-norm 2013: Normalización léxica de tuits en español, 2013.

[4] Apache. Opennlp. http://opennlp.apache.org.

[5] AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. A phrase-based statistical model for sms text normalization, 2009.

[6] Richard Beaufort, Sophie Roekhaut, Louise-Amelie Cougnon, and Cedrick Fairon. A hybrid rule/model-based finite-state framework for normalizing sms messages, 2002.

[7] Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model, 2003.

[8] Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction, 2000.

[9] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class based n-gram models of natural language, 1992.

[10] Jhon Adrián Cerón-Guzmán and Elizabeth León-Guzmán. Lexical normalization of spanish tweets, 2016.

[11] Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. Investigation and modeling of the structure of texting language, 2007.

[12] Paul Cook and Suzanne Stevenson. An unsupervised model for text message normalization, 2009.

[13] J.M. Cotelo, F.L. Cruz, J.A. Troyano, and F.J. Ortega. A modular approach for lexical normalization applied to spanish tweets, 2015.

[14] dwyl. english-words. https://github.com/dwyl/english-words.

[15] dylon. liblevenshtein. https://github.com/universal-automata/liblevenshtein-java.

[16] Jacob Eisenstein. What to do about bad language on the internet, 2013.

[17] Facebook. fasttext. https://fasttext.cc/.

[18] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in twitter data with crowd-sourcing, 2010.

[19] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling, 2005.

[20] Jenny Rose Finkel and Christopher D. Manning. Nested named entity recognition, 2009.

[21] Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. From news to comment: Resources and benchmarks for parsing the language of web 2.0, 2011.

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

[23] Pablo Gamallo, Marcos García, and Santiago Fernández-Lanza. Word normalization in twitter using finite-state transducers, 2013.

[24] Pablo Gamallo, Marcos García, and José Ramon Pichel. A method to lexical normalisation of tweets, 2013.

[25] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael, Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: annotation, features, and experiments, 2011.

[26] Jose M. Gomez-Hidalgo, Andrés A. Caurcel-Díaz, and Yovan Iñiguez del Rio. Un método de análisis de lenguaje tipo sms para el castellano, 2013.

[27] Google. Google cloud engine. https://cloud.google.com/.

[28] Google. Word2vec. https://github.com/deeplearning4j/deeplearning4j.

[29] Stanford NLP Group. Stanfordnlp core. https://stanfordnlp.github.io/CoreNLP/.

[30] Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Xian Wu, and Zhong Su. Domain adaptation with latent semantic association for named entity recognition, 2009.

[31] Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a twitter, 2011.

[32] Bo han, Paul Cook, and Timothy Baldwin. unimelb: Spanish text normalisation, 2013.

[33] K. Heafield. Faster and smaller language model queries, 2011.

[34] Mans Hulden and Jerid Francom. Weighted and unweighted transducers for tweet normalization, 2013.

[35] Martin Jansche and Steven P. Abney. Information extraction from voicemail transcripts, 2002.

[36] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp., 2007.

[37] Joseph Kaufmann and Jugal Kalita. Syntactic normalization of twitter messages, 2010.

[38] Catherine Kobus, Franois Yvon, and Graldine Damnati. Transcrire les sms comme on reconnat la parole, 2008.

[39] George R. Krupka and Kevin Hausman. Isoquest: Description of the netowltm extractor system as used in muc-7., 1998.

[40] Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets, 2011.

[41] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Cernocký. Empirical evaluation and combination of advanced language modeling techniques, 2011.

[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[43] Scott Miller, Jethran Guinness, and Alex Zamanian. Name tagging with word clusters and discriminative training, 2004.

[44] Einat Minkov, Richard C. Wang, and William W. Cohen. Extracting personal names from email: applying named entity recognition to informal text, 2005.

[45] Javier Moreno. Tweetsc spell checker app. https://jmorenov.github.io/TweetSC/.

[46] Javier Moreno. Tweetsc web. https://tweetsc.github.io.

[47] Mosquera, Alejandro, Elena Lloret, and Paloma Moreda. Towards facilitating the accessibility of web 2.0 texts through text normalisation, 2012.

[48] Alejandro Mosquera. Algoritmo del metáfono. https://github.com/amsqr/Spanish-Metaphone/blob/master/phonetic_algorithms_es.py.

[49] Alejandro Mosquera and Paloma Moreda. Dlsi en tweet-norm 2013: Normalizacion de tweets en español, 2013.

[50] Juan M. Cotelo Moya, Fermín L Cruz, and Jose A. Troyano. Resource-based lexical approach to tweet-norm task, 2013.

[51] Forsyth Eric N. and Craig H. Martell. Lexical and discourse analysis of online chat dialog, 2007.

[52] Peter Norvig. How to write a spelling corrector. http://norvig.com/spell-correct.html, 2007.

[53] Jesús Oliva, José I. Serrano, María D. Del Castillo, and Angel Iglesias. Sms normalization: combining phonetics, morphology and semantics, 2011.

[54] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters, 2013.

[55] James L. Peterson. Computer programs for detecting and correcting spelling errors., 1980.

[56] Bojanowski Piotr, Grave Edouard, Joulin Armand, and Mikolov Tomas. Enriching word vectors with subword information, 2017.

[57] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations, 2010.

[58] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: an experimental study, 2011.

[59] Pablo Ruiz, Montse Cuadros, and Thierry Etchegoyhen. Lexical normalization of spanish tweets with preprocessing rules, domain-specific edit distances, and language models, 2013.

[60] Pablo Ruiz, Montse Cuadros, and Thierry Etchegoyhen. Lexical normalization of spanish tweets with rule-based components and language models, 2014.

[61] Arturo Montejo Ráez, M. Carlos Diaz Galiano, Eugenio Martíne Cámara, M. Teresa Martín Valdivia, Miguel A. García Cumbreras, and L. Alfonso Ureña López. Sinai at twitter-normalization 2013, 2013.

[62] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: language independent named entity recognition, 2003.

[63] H. Schwenk. Continuous space language models, 2007.

[64] Claude Elwood Shannon. A mathematical theory of communication, 1948.

[65] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically, 2010.

[66] Sameer Singh, Dustin Hillard, and Chris Leggetter. Minimally-supervised extraction of entities from text advertisements, 2010.

[67] Enrique Sánchez-Villamil, Mikel L. Forcada, and Rafael C. Carrasco. Unsupervised training of a finite-state sliding-window part-of-speech tagger, 2004.

[68] Kristina Toutanova and Robert C. Moore. Pronunciation modeling for improved spelling correction, 2002.

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

[70] Xabier Saralegi Urizar and Iñaki San Vicente Roncal. Elhuyar at tweetnorm 2013, 2013.

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

[72] Jesús Vilares, Miguel A. Alonso, and David Vilares. Prototipado rápido de un sistema de normalización de tuits: Una aproximación léxica, 2013.

[73] Casey Whitelaw, BenHutchinson, Grace Y. Chung, and Gerard Ellis. Using the web for language independent spellchecking and autocorrection, 2009.

[74] Dan Wu, Wee Sun Lee, Nan Ye, and Hai Leong Chieu. Domain adaptive bootstrapping for named entity recognition, 2009.

[75] GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger., 2002.

[76] Óscar Muñoz-García, Silvia Vázquez, and Nuria Bel. Exploiting web-based collective knowledge for micropost normalisation, 2013.