Int. al Procesamiento de Lenguaje Natural
Laboratorio 2017

El objetivo del laboratorio 2017 del curso Introducción al Procesamiento de Lenguaje Natural es identificar textos humorísticos en tweets en español. Para esto, el estudiante deberá familiarizarse con diferentes herramientas de PLN de acceso libre e incorporarlas en un entorno unificado de programación (en este caso, sobre la plataforma Python).

Corpus

El corpus a utilizar consta de tweets en idioma español, donde una gran parte de los tweets son chistes o algún otro tipo de texto humorístico. El corpus contiene anotaciones hechas por usuarios humanos (anotadores) acerca de su opinión sobre cada tweet: si el contenido del tweet es humorístico o no, y en caso de ser humorístico qué tan gracioso es en su opinión (en una escala de 1 a 5 estrellas).

El corpus está compuesto por 15134 tweets y tiene la siguiente estructura:

- **id:** Identificador numérico único del tweet.
- **text**: Texto completo del tweet.
- **account_id:** Identificador numérico único de la cuenta de Twitter que publicó el tweet.
- **n:** Cantidad de votos que opinan que el contenido del tweet no es humorístico.
- **1:** Cantidad de votos que opinan que el tweet es humorístico y lo califican con 1 estrella.
- **2:** Igual al anterior, pero lo califican con 2 estrellas.
- **3:** Igual al anterior, pero lo califican con 3 estrellas.
- **4:** Igual al anterior, pero lo califican con 4 estrellas.
- **5:** Igual al anterior, pero lo califican con 5 estrellas.

La siguiente es una de las instancias:

```
24282566299230208,"Borracho en est. policía: -Podría ver al q robó anoche en mi casa? -Para qué? -Para saber cómo entró sin despertar a mi mujer! by @edwinm53",132679073,0,0,0,1,1,0
```

En este caso, el chiste fue calificado como de 4 estrellas por un anotador, y como de 3 estrellas por otro.

Descripción del trabajo

El laboratorio consta de tres tareas principales:

i) Construir un clasificador binario que determine si un tweet dado es humorístico o no. Para ello, se considerará "humorístico" si la mitad o más de los anotadores lo calificaron con una o más estrellas, y "no humorístico" en caso contrario.

ii) Construir un clasificador que, además de determinar si el tweet es humorístico o no, prediga la mediana de la cantidad de estrellas asignadas por los anotadores (contando como 0 los votos por "No Humorístico". Por ejemplo, si un tweet tiene un voto como "No Humor", dos votos para dos estrellas y dos votos para tres estrellas, la mediana vale 2.

iii) A partir del resultado del paso ii, y considerando como "no humorístico" a todos aquellos tweets cuya mediana sea menor a 1, construir un clasificador como en el paso i) y comparar los resultados

Al comienzo del laboratorio, se dejará disponible el corpus para realizar el entrenamiento, así como un notebook Jupyter con los comandos para importarlo. Diez días antes del plazo de entrega, se liberará un corpus de evaluación donde deberán ejecutarse los clasificadores construidos, y evaluar la performance.

Atención: para construir los clasificadores y evaluarlos solamente se deberán tomar en cuenta aquellos tweets que tienen, al menos, 3 votos. Es responsabilidad del alumno filtrarlos antes de entrenar/evaluar. Esto deberá quedar claramente documentado.

Metodología

Los estudiantes deberán definir una metodología para el entrenamiento y la evaluación, de acuerdo a lo visto en el curso. Esto incluye la definición de los corpus de entrenamiento, validación y evaluación, y las medidas de evaluación a utilizar .

Herramientas

- Se utilizará la plataforma Python para el desarrollo, en su versión 3.6 o superior. En particular, los resultados se probarán sobre una distribución Anaconda, por lo que se sugiere instalarla.
- La biblioteca para procesamiento de lenguaje natural sobre Python a utilizar es NLTK
- Para la tokenización, tagging y cualquier tarea para el idioma español, se tiene la herramienta FreeLing, la que deberá integrarse al ambiente.
- Podrán utilizarse bibliotecas adicionales de Python, tales como: NumPy y SciPy, pandas o scikit-learn.

Podrán utilizarse herramientas y bibliotecas adicionales, siempre y cuando se integren al entorno Python de trabajo.

Formato de entrega

La entrega deberá realizarse utilizando un notebook Jupyter donde se incluirá tanto la documentación como el código Python a ejecutar.

Las tareas a realizar y documentar en el notebook son las siguientes:

- Importación del corpus de entrenamiento.
- Análisis del corpus: deberán definirse propiedades y calcularse estadísticas de utilidad para describir el corpus, tanto desde un punto de vista cuantitativo como cualitativo.

- Preprocesamiento: deberán eliminarse los hashtags de los tweets (dado que ello facilitaría demasiado la clasificación).
- Clasificación: deberán entrenarse clasificadores sobre el corpus según lo especificado. Para esto, se deberán utilizar como atributos, al menos, las palabras y los POS tags resultantes del análisis con Freeling, y todos los otros atributos que considere convenientes.
- Descripción de los resultados en el corpus de entrenamiento o validación, con las medidas definidas, e incluyendo las correspondientes matrices de confusión.
- Descripción de los resultados en el corpus de evaluación.

El notebook deberá incluir una descripción clara y justificada del proceso realizado, un análisis cualitativo y cuantitativo de los resultados obtenidos, y sugerencias de posibles mejoras a futuro.

Evaluación

Para la evaluación del laboratorio, se tendrá en cuenta:

- El cumplimiento de las tareas pedidas, incluyendo el código solicitado.
- La claridad de la documentación, en particular la justificación de las decisiones tomadas, así como el análisis de los resultados obtenidos.

Deberá presentarse en la entrega toda la información de configuración necesaria para poder reproducir el proceso. Esto será condición necesaria para la aprobación del curso.

Insumos

Se proveerán los siguientes archivos para la realización del laboratorio:

- corpus_humor_training.csv: corpus de tweets en español etiquetado con información referente al humor. Conjunto de entrenamiento y desarrollo.
- corpus_humor_testing.csv: corpus de tweets en español etiquetado con información referente al humor. Conjunto de evaluación.
- Laboratorio_IntroPLN2017.ipynb: notebook Python con ejemplo de carga de datos.

Referencias

Castro y Cubero: Detección de humor en textos en español
Bird, Klein, Lopper: “Natural Language Processing with Python”
Referencia de NLTK
The Jupyter notebook
Freeling 4.0, sitio oficial