Filtrado de correo-e no deseado

Inteligencia Artificial – Ingeniería del Software
Curso 2021/2022
Propuesta de trabajo

Álvaro Romero Jiménez

1. Introducción y objetivos

En la comunicación por correo electrónico (correo-e, de aquí en adelante) se emplea el término *spam* para designar a aquellos mensajes enviados de forma masiva y que no han sido solicitados por sus destinatarios. Los gestores de correo-e suelen incorporar la posibilidad de crear filtros de mensajes no deseados que permitan identificar estos de forma automática y actuar en consecuencia (usualmente moviéndolos de la bandeja de entrada a una carpeta específica). El aumento del volumen de correos electrónicos no deseados recibidos por los usuarios ha creado la necesidad de desarrollar filtros más fiables y robustos, para lo que se aplica toda una serie de métodos, entre los que tienen una gran prevalencia aquellos provenientes del campo del aprendizaje automático.

El **objetivo principal** de esta propuesta es construir, mediante técnicas de procesamiento del lenguaje natural, un filtro de correo-e no deseado que resuelva esa tarea de la mejor manera posible.

Para ello será necesario alcanzar los siguientes **objetivos específicos**:

1. Recopilar distintas técnicas de procesamiento del lenguaje natural aplicables a la tarea a realizar.
2. Construir filtros de correo-e no deseado que incorporen las técnicas consideradas.
3. Evaluar los filtros construidos, escogiendo para ello las métricas que se estimen más adecuadas.
4. Seleccionar, de entre todos los filtros construidos, el que se juzgue que realiza mejor la tarea.
5. Documentar el trabajo realizado usando un formato de artículo científico.
6. Realizar una presentación (PDF, PowerPoint o similar) de los resultados obtenidos.

2. Descripción del trabajo

A continuación se describe con más detalle cómo debe llevarse a cabo el trabajo.

Metodología

Enron-Spam es un conjunto, que está disponible públicamente, de mensajes (en inglés) de correo-e, ya preclasificados en mensajes legítimos y no deseados. Para la realización de este trabajo se proporciona un subconjunto de los mismos, reservándose una parte para la evaluación del trabajo, tal y como se explica en la sección 3. Los mensajes se proporcionan en bruto, incluyendo las cabeceras (para su lectura se recomienda usar el paquete email, incluido en la biblioteca estándar de Python).

Inicialmente, se pide construir dos tipos de filtros de la siguiente manera (se recomienda el uso de los paquetes scikit-learn y NLTK de Python):

- Seleccionar un vocabulario de términos adecuado, a partir del análisis de los mensajes proporcionados.
- Construir un filtro usando la bolsa de palabras como modelo de lenguaje y naive Bayes multinomial como modelo clasificador (experimentar con distintos valores del hiperparámetro de suavizado).
- Construir un filtro usando tf-idf como modelo de lenguaje y kNN como modelo clasificador (experimentar con distintos valores del hiperparámetro k).

A partir de aquí, se pide mejorar los filtros obtenidos incorporando alguna o varias de las siguientes técnicas:

- Técnicas de preprocesamiento de datos: eliminación de ruido, tokenización, normalización.
- Uso de atributos derivados de los mensajes, tales como: indicación de que el mensaje es una respuesta o un reenvío; tamaño del asunto y/o cuerpo del mensaje; presencia de etiquetas HTML en el cuerpo del mensaje; etc. Una tabla bastante completa de posibles atributos a considerar se encuentra en el artículo *Applicability of machine learning in spam and phishing email filtering: review and approaches*.
- Cualquier otra técnica que se considere útil.

Para evaluar el rendimiento de los filtros construidos se debe, en primer lugar, dividir el conjunto de mensajes proporcionado en dos subconjuntos, uno de entrenamiento y el otro de validación. Para ello puede resultar útil la biblioteca split-folders de Python.

Se debe proceder, entonces, a entrenar los distintos filtros a partir del subconjunto de entrenamiento y estimar su rendimiento sobre el subconjunto de validación. Para esto último deben seleccionarse, justificadamente, una o varias métricas de clasificación binaria que permitan comparar los filtros entre sí.

De entre todos los filtros construidos, se deberá finalmente escoger uno para ser evaluado sobre el subconjunto de prueba reservado aparte por el profesor.

Documentación y entrega

El trabajo deberá documentarse siguiendo un formato de artículo científico, con una **extensión mínima de 6 páginas**. En la página web de la asignatura se pueden encontrar plantillas donde se sugiere una estructura general. Estas plantillas siguen el formato de

los *IEEE conference proceedings*, cuyo sitio web guía para autores ofrece información más detallada. El documento entregado deberá estar en formato PDF. Se valorará el uso del sistema LaTeX.

En el caso concreto de este trabajo, la memoria deberá al menos incluir: introducción; descripción de los filtros construidos y de las técnicas que incorporan; descripción de los resultados alcanzados; conclusiones; bibliografía. **En ningún caso debe incluirse código en la memoria**.

La entrega del trabajo consistirá de **un único fichero comprimido zip** conteniendo la memoria, el código implementado (ficheros `py` o cuadernos de Jupyter) y un fichero `py` (no cuaderno de Jupyter) que debe llamarse `clasificador_seleccionado.py` y debe incluir una función llamada `es_mensaje_no_deseado` que, al recibir la ruta a un fichero conteniendo un mensaje de correo-e, debe devolver `True` si se trata de un mensaje no deseado, según el filtro seleccionado, y `False` en caso contrario. Esta función es la que se usará para evaluar el filtro sobre el conjunto de prueba reservado por el profesor. **En caso de no proporcionar un fichero con ese nombre exacto o que no defina la función con ese nombre exacto, esa parte de la evaluación no se realizará**.

Presentación y defensa

Como parte de la evaluación del trabajo se deberá realizar una defensa del mismo, para lo que se citará a los alumnos de manera conveniente.

El día de la defensa se deberá realizar una pequeña presentación (PDF, PowerPoint o similar) de 10 minutos en la que participarán activamente todos los miembros del grupo que ha desarrollado el trabajo. Esta presentación deberá seguir a grandes rasgos la misma estructura que la memoria del trabajo, haciendo especial mención a los resultados obtenidos y al análisis crítico de los mismos.

En los siguientes 10 minutos de la defensa, el profesor procederá a realizar preguntas sobre el trabajo, que podrán ser tanto de la memoria como del código fuente.

3. Evaluación del trabajo

Para la evaluación del trabajo se tendrán en cuenta los siguientes criterios, considerando una nota total máxima de 4 puntos:

- *Memoria del trabajo* (hasta 1 punto): se valorará la claridad de las explicaciones, el razonamiento de las decisiones, el análisis y presentación de resultados y el correcto uso del lenguaje. La elaboración de la memoria debe ser original, por lo que no se evaluará el trabajo si se detecta cualquier copia del contenido.

- *Código fuente* (hasta 1 punto): se valorará la claridad y buen estilo de programación, corrección y eficiencia de la implementación y calidad de los comentarios. El código debe ser original, por lo que no se evaluará el trabajo si se detecta código copiado o descargado de internet.

- *Filtro seleccionado* (hasta 1 punto): se valorará, tanto de manera absoluta como comparativamente con el resto de trabajos, el comportamiento del filtro seleccionado sobre el conjunto de mensajes de prueba reservado.

- *Presentación y defensa* (hasta 1 punto): se valorará la claridad de la presentación y la buena explicación de los contenidos del trabajo así como, especialmente, las respuestas a las preguntas realizadas por el profesor.

IMPORTANTE: cualquier plagio, compartición de código o uso de material que no sea original y del que no se cite convenientemente la fuente, significará automáticamente la calificación de cero en la asignatura para todos los alumnos involucrados. Por tanto, a estos alumnos no se les conserva, ni para la actual ni para futuras convocatorias, ninguna nota que hubiesen obtenido hasta el momento. Todo ello sin perjuicio de las correspondientes medidas disciplinarias que se pudieran tomar.