Clasificación de documentos con word2vec

Objetivo

Utilizar técnicas de aprendizaje supervisado o no supervisado para clasificar fragmentos de texto en función de su temática, empleando una vectorización numérica de la información de los textos.

Desarrollo de la práctica

A continuación, se exponen los pasos necesarios para la realización del trabajo:

1. Construcción del corpus. Para esto se deben buscar fragmentos (en español o en inglés) de distintos artículos periodísticos (u otro tipo de documentos) y etiquetarlos según sus temáticas, como por ejemplo: naturaleza, finanzas, arquitectura, … El tamaño del corpus debe ser adecuado para las tareas posteriores, se recomienda no empezar por corpus muy extensos y aumentarlo en función de los resultados. La elección de las posibles temáticas es libre, y será la clase que se intentará predecir.

2. Crear una función que tokenice los textos y elimine las stop words (o palabras que no portan información sobre el problema). En este paso se limpian los textos para convertirlos en vectores de palabras que porten información y se eliminan mayúsculas y símbolos de puntuación. Por ejemplo, en la clasificación de documentos se espera una mayor capacidad predictora de la clase para la palabra “tigre” que para la palabra “pero”. Las listas de stop words de cada idioma pueden descargarse directamente de Internet.

3. Emplear la librería NLTK para aplicar stemming sobre el corpus. Por último, crear el vocabulario final en función de la frecuencia de las palabras que aparezcan en el corpus (pudiendo no considerar las de baja frecuencia según un umbral libre). Alternativamente, a la hora de crear el modelo en el próximo paso existen argumentos para realizar este filtrado en base a conteos.

4. Consultar la documentación de word2vec, explicar de forma breve en qué se basa y entrenar un modelo de este tipo sobre el conjunto limpio para vectorizar las frases usando la librería “gensim”. En este punto, ilustrar con algún ejemplo el parecido entre algunas palabras en base a los valores de sus embeddings.

En función de si la tarea se va a presentar en junio, julio o noviembre, las alternativas a partir de este punto son:

JUNIO. Clasificación. A partir de los resultados del embedding, entrenar algunos modelos que permitan clasificar el tipo de documentos a los que pertenecen las frases en función de las palabras que aparecen y comentar los resultados de forma apropiada. Entrenar después un modelo Naive Bayes multinomial sobre la bolsa de palabras creada antes de aplicar word2vec y comparar resultados. Para esto, utilizar el flujo de trabajo típico de la librería scikit-learn.

JULIO Y NOVIEMBRE. Clustering. A partir de los resultados del embedding, utilizar distintas técnicas de clustering para agrupar los documentos, creando representaciones gráficas y discutiendo si los grupos creados guardan relación con las etiquetas originales. Evaluar en función de métricas apropiadas.

Documentación y entrega

El trabajo deberá documentarse siguiendo un formato de artículo científico correspondiente a IEEE conference proceedings, cuyo sitio web guía para autores ofrece información detallada. El documento entregado deberá estar en formato PDF. La memoria deberá incluir: introducción al problema de clasificación de documentos y aplicaciones, funcionamiento de word2vec, justificación de cada apartado del desarrollo, resultados debidamente discutidos y conclusiones del trabajo. Será necesario incluir las referencias consultadas. El código empleado en la resolución del problema se entregará como notebook de Jupyter. Para su entrega, ambos ficheros se subirán a la plataforma en una carpeta comprimida en formato zip.

Criterios de evaluación

El trabajo se evaluará de acuerdo a los siguientes criterios:

- Memoria escrita (1,5 puntos). Se valorará la claridad de las explicaciones, el razonamiento de las decisiones, el análisis y presentación de resultados y el correcto uso del lenguaje.
- Código utilizado para la resolución del problema (1,5 puntos). Se valorará la claridad, corrección y eficiencia de la implementación. Se deberán incluir comentarios que faciliten la interpretabilidad del código empleado.
- Presentación y defensa (1 punto). Se valorará la claridad de la presentación y la buena explicación de los contenidos del trabajo, así como las respuestas a las preguntas realizadas por el profesor.

La detección de plagio en el contenido de la memoria o del código implicará la adopción de las medidas postuladas en el proyecto docente de la asignatura.