Tarea del Tema 5: Estudio de diferentes representaciones vectoriales

El objetivo de esta tarea es crear representaciones de documentos dentro del *modelo de espacio vectorial* y de *modelos semánticos vectoriales*.

En primer lugar, deberemos generar cuatro pequeñas colecciones de documentos descargados a tu elección, y formadas por páginas web en inglés y español, y por tweets escritos en inglés y español.

Para generar las representaciones dentro del modelo de espacio vectorial deberás aplicar las siguientes fases por cada una de las colecciones:

- **Análisis léxico**; seleccionando los rasgos con los que generar las representaciones. Deberán considerarse los espacios en blanco como separadores entre cadenas. El resto de decisiones como la eliminación de signos de puntuación, caracteres raros, etc. se dejan a la elección del estudiante.
- Eliminación de **stop-words**. Deberá buscarse una lista de stop-words en inglés y otra en español, y aplicarlas durante la fase de eliminación de palabras vacías.
- **Truncado**; aplicando el algoritmo de *stemming* correspondiente:
  - Inglés: http://tartarus.org/martin/PorterStemmer/
  - Español: http://snowball.tartarus.org/algorithms/spanish/stemmer.html
- Aplicación de las **funciones de pesado** TF y TF-IDF.

En el caso de los modelos de representación semántica, deberás partir de *word embedings* preentrenados por idioma:

- Inglés: https://code.google.com/archive/p/word2vec/
- Español: https://crscardellino.github.io/SBWCE/

Y para generar las representaciones de documentos aplicaremos dos de los baselines más utilizados dentro de la representación semántica distribucional, el modelo aditivo (*additive model*), que compone la representación de una frase sumando los vectores de representación v_i de las palabras contenidas en la misma:

$$f(v_1, v_2) = v_1 + v_2$$

y el de la media (*average model*), que compone la representación haciendo la media de los vectores de la composición:

$$f(v_1, v_2) = \frac{v_1 + v_2}{2}$$

> **Nota**: Debe tenerse en cuenta que, a partir de la definición anterior del *average model*, para hacer la composición en frases con más de dos palabras habrá que decidir el orden en el que ir componiendo la palabras de la frase de dos a dos; esto es porque en una frase con tres términos, el resultado de aplicar `f(f(v1,v2),v3)` no tiene por qué ser igual que aplicar el orden `f(v1,f(v2,v3))` si `f()` es la función correspondiente al *average model*. En el caso del *additive model* el orden no influye en el resultado final, ya que el resultado de sumar en diferente orden los vectores de las palabras de una frase siempre será el mismo.
>
> Se deja a la elección del estudiante el orden de aplicación de la función de composición *average model*, siempre que se detalle en la memoria y se pueda comprobar que según el orden establecido, el *embedding* de la frase se genera de manera correcta.

Un "orden" típico sería el orden de lectura, es decir, de izquierda a derecha, de modo que una fase “v1 v2 v3” se compondría como `f(f(v1,v2),v3)`. Otra opción podría ser en sentido inverso, de derecha izquierda, correspondiéndose .con la expresión `f(v1,f(v2,v3))`.

Deberán entregarse ficheros de texto con las representaciones generadas.

Finalmente se deberá entregar una pequeña memoria que contenga una sección para cada uno de los modelos de representación, destacando los aspectos considerados en cada caso y donde se expliquen razonadamente las decisiones que se han ido tomando.