Práctica de Modelado de Tópicos con Latent Dirichlet Allocation

Objetivo

La cuarta y última práctica de la asignatura de procesamiento de lenguaje natural y minería de textos engloba conocimientos y habilidades adquiridas durante de las prácticas y sesiones mantenidas hasta ahora. El objetivo es realizar una extracción de tópicos de una serie de documentos, para lo cual será necesario realizar el pre-procesamiento completo de los mismos. Se podrán utilizar todas las librerías utilizadas con anterioridad, como NLTK o spaCy, si bien los textos estarán en castellano, y no todas las librerías presentan las mismas funcionalidades para nuestro idioma.

1. Importe un conjunto de al menos 10 documentos. Los documentos pueden importarse de ficheros que estén en una carpeta del sistema operativo, o bien de cadenas de textos que introduzca en el código de la práctica. Deben estar en castellano, y deben tener un contenido suficientemente rico para la práctica, como tener entidades a extraer, al menos de tipo lugar, persona y organización.
2. Los textos deben cubrir varios tópicos, a elección del alumno. Por ejemplo, pueden hablar de deportes, cultura, alimentación, flora y fauna, etc. No es preciso que sean muy largos, no más de 3 ó 4 frases por documento
3. Haga un análisis completo de cada texto. Divídalo en frases, palabras, haga el análisis gramatical (POS) de cada palabra, saque las formas normales y extraiga entidades. Liste únicamente los valores (frases, palabras, POS, forma normal y entidades) del primero de los documentos. Se recomienda utilizar la librería spaCy
4. Prepare el conjunto de documentos para realizar un análisis de tipo LDA. Será necesario pasar todo a minúsculas, pasar a formas normales y retirar palabras vacías.

5. De cara a realizar el análisis LDA tendrá que obtener el diccionario de los términos de los documentos, así como la matriz de términos documentos.
6. Genere el modelo LDA y realice el análisis. Se recomienda utilizar la librería gensim
7. Por último, represente los valores obtenidos. Al menos represente cada tópico con los 5 términos más frecuentes en cada uno de ellos.
8. Intente obtener el resultado gráfico del análisis con la librería pyLDAvis. Esta librería crea un modelo gráfico interactivo con los datos de LDA, y podremos ver gráficamente los clusters generados, así como la distancia entre los mismos y las palabras más frecuentes
9. Entregue un notebook con el programa y sus resultados. Si ha utilizado ficheros externos para los documentos, entregue también un fichero zip con los documentos

Nota.- La librería pyLDAvis tiene que instalarse previamente, y puede ser algo complicado, y dar errores en tiempo de ejecución del código. La mayor parte de las veces, estos errores se deben a librerías que no están cargadas, o a incompatibilidad con versiones antiguas. Se recomienda.

Cuando se recibe un mensaje de error como por ejemplo:

No module named …….. in matplotlib

Es posible que la librería en cuestión (matplotlib) no esté cargada. pyLDAvis puede depender de ella, pero no se ha cargado al instalar pyLDAvis. En este caso, instalar matpotlib. Por el contrario, si ya estuviese cargada, es posible que sea necesario actualizarla. Para ello, si se utiliza Anaconda, se puede emplear la instrucción

Conda update –all -y