Laboratorio 5.
Predicción de nuevas palabras usando minería de texto.

INSTRUCCIONES:

Utilice el data set Tweets Blogs News - Swiftkey Dataset 4million. NLP - Tweets, Blogs, and News Articles 4 million text entries de Kaggle. Debe hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final es predecir la próxima palabra que el usuario ingresará. El sistema debe proponer las próximas 3 palabras más probables, que el usuario escribirá, dada una frase que ingrese. Genere un informe en pdf con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios y/o cada uno de los pasos llevados a cabo si utiliza una herramienta visual. Incluya una nube de palabras que le ayude a detectar las que más se repiten. Este laboratorio debe realizarse en grupos de 3. Inscríbase en uno de los grupos que hay en canvas para la actividad.

DESCRIPCIÓN DEL DATASET

Los datos consisten en archivos que contienen textos de noticias, tweets, y blogs en 4 idiomas:

- Inglés
- Alemán
- Ruso
- Francés

Utilice los archivos del idioma Inglés para hacer sus predicciones

EJERCICIOS

1. Descargue los archivos de texto en inglés
2. Cargue los archivos de datos a R o a Python, dependiendo de con qué trabaje.
3. Limpie y preprocese los datos. Describa de forma detallada las actividades de preprocesamiento que llevó a cabo.
   3.1. Se pueden hacer tareas como:
   - Convertir el texto a mayúsculas o a minúsculas
   - Quitar los caracteres especiales que aparecen como “#”,”@” o los apóstrofes.
   - Quitar las url
   - Revisar si hay emoticones y quitarlos
   - Quitar los signos de puntuación
   - Quitar los artículos, preposiciones y conjunciones (stopwords)
   - Quitar números si considera que interferirán en las predicciones.
4. Obtenga la frecuencia de las palabras de cada archivo.
5. Haga un análisis exploratorio de los datos para entenderlos mejor, documente todos los análisis
   5.1. Puede, para cada archivo:
   - Investigar qué palabra se repite más en cada archivo
   - Hacer una nube de palabras para visualizar las que aparecen con más frecuencia
   - Hacer un histograma con las palabras que más se repiten

- Discutir sobre las palabras que tienen presencia en todos los archivos.

6. Utilice una muestra de los archivos para poder obtener los n-gramas (Feinerer, Hornik, & Meyer, 2008; Jurafsky & Martin, 2014). Un 5% de cada archivo puede funcionar. Mientras más grande la muestra, mejor serán las predicciones.
7. Obtenga las matrices de términos para cada palabra, 2-grama (combinación de dos palabras) o 3-grama (combinación de tres palabras).
8. Calcule la probabilidad de ocurrencia de cada uno de los n-gramas, puede utilizar Kneser-Ney smoothing probability (KNP), para predecir las siguientes 3 palabras. Guarde esos datos.
9. Elabore una función en la que el usuario ingrese una frase y el sistema prediga (Daniel Jurafsky, 101AD) las próximas tres palabras posibles, con su respectivo valor de probabilidad. En el siguiente vínculo tiene un ejemplo de lo que debería hacer su función. https://lgarciap.shinyapps.io/TextPredictor/

EVALUACIÓN

(25 puntos) Análisis exploratorio:
- Se elaboró un análisis exploratorio en el que se explican los cruces de variables, hay gráficos explicativos y análisis que permiten comprender el conjunto de datos.

(20 puntos) Limpieza y preprocesamiento de los datos:
- Se documentan las tareas de limpieza, incluyendo los paquetes/módulos que se usaron.

(20 puntos) Generación de los ngramas y cálculo de sus frecuencias y probabilidades:
- Se explica como se generaron los n-gramas y se calcularon los valores de frecuencia y probabilidades.

(10 puntos) Algoritmo:
- Se describe el algoritmo que se usó para predecir.

(25 puntos) Función de predicción.
- Se elaboró una función que permite predecir las n posibles palabras que escribirá el usuario tras la frase ingresada.

MATERIAL A ENTREGAR

- Archivo .pdf con el informe que contenga, los resultados de los análisis y las explicaciones.
- Link de Google drive donde trabajó el grupo.
- Script de R (.r o .rmd) o de Python que utilizó.
- Link del repositorio usado para versionar el código.

FECHAS DE ENTREGA

- **AVANCE:** Descripción de los datos (tamaño, origen y tipo de corpus), preprocesamiento y sus explicaciones, unigramas, bigramas, modelo preliminar de predicción: miércoles 25 de septiembre 11:18.
- **DOCUMENTO FINAL COMPLETO:** martes 1 de octubre de 2019 a las 23:59
  NOTA: Solo se calificará el Documento Final si está entregado el avance con todo lo que se pide.

REFERENCIAS

Daniel Jurafsky, J. H. M. (101AD). Speech and Language Processing (2008), *1*. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal Of Statistical Software*, *25*(5), 1–54. https://www.jstatsoft.org/article/view/v025i05

Jurafsky, D., & Martin, J. H. (2014). N-Grams. *Speech and Language Processing*, 2–7. Retrieved from https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf

PAQUETES ÚTILES DE R

- Quanteda
- Wordcloud
- Tm
- Rweka
- Ngram