Inteligencia artificial avanzada para la ciencia de datos I

TC3006C

Dr. Esteban Castillo Juarez
TEC de Monterrey, Campus Santa Fe


esteban.castillojz@tec.mx

Tecnológico de Monterrey

Agenda

- Objetivo
- Instrucciones
- Conjunto de datos de X
- Formato de los vectores
- Formato de la actividad

Tecnológico de Monterrey

Objetivo

- Reforzar la teoría de Naive Bayes aplicando el algoritmo de clasificación sobre el mismo conjunto de datos utilizado en las sesiones, pero con más clases (etiquetas de clasificación).

- Practica la teoría detrás de Naive Bayes en el contexto del lenguaje de programación Python con Scikit-learn.

- Reforzar tópicos como el uso de métricas o la validación cruzada.

Tecnológico de Monterrey

Instrucciones

Para la implementación manual y con Scikit-learn hacer lo siguiente:

1. Utilizar el conjunto de datos de la plataforma X usado en la sesión de clase.

2. Adaptar el código de Naive Bayes para manejar más de dos clases, considerando que la nueva versión del conjunto de datos de X incluye muestras positivas, negativas y **neutrales**.

3. Modificar la forma en que se crean los vectores utilizando el enfoque de '`Bag of Words`' en lugar de '`One-Hot-Encoding`'.

Tecnológico de Monterrey

Instrucciones

Para la implementación manual y con Scikit-learn hacer lo siguiente:

4. Entrena el modelo utilizando `20, 40, 60, 80, 100` y `120` características de tu preferencia (vocabulario). Evalúa el desempeño aplicando validación cruzada con `K = 3, 4, 5` y 6, y emplea las métricas de precisión, recuerdo y F1.

5. Proponga seis visualizaciones que ayuden a entender como se comportan los modelos entre distintas iteraciones.

6. Describa los hallazgos y proporcione una conclusión sobre el proceso de experimentación ¿Qué implementación fue mejor? ¿Qué características ayudaron mas? Etc.

Conjunto de datos de X

- El conjunto de datos es una variación del utilizado en la sesión de clase. Este conjunto de datos condensa muestras de mensajes etiquetados en tres categorías principales: positivo, negativo y neutral.

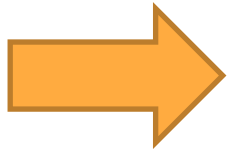
- Este conjunto de datos fue preprocesado para evitar problemas de codificación. El preprocesamiento incluyó la eliminación de símbolos de puntuación, URLs, espacios extra y elementos que no forman parte de la codificación ASCII.

Tecnológico de Monterrey

Conjunto de datos de X

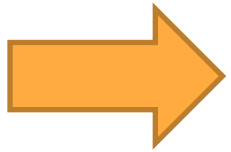
Para esta nueva variación se tiene la siguiente distribución:

- Entrenamiento:
  - 2249 muestras positivas
  - 859 muestras negativas
  - 1079 muestras neutrales



4187 muestras

- Prueba
  - 358 muestras positivas
  - 179 muestras negativas
  - 330 muestras neutrales

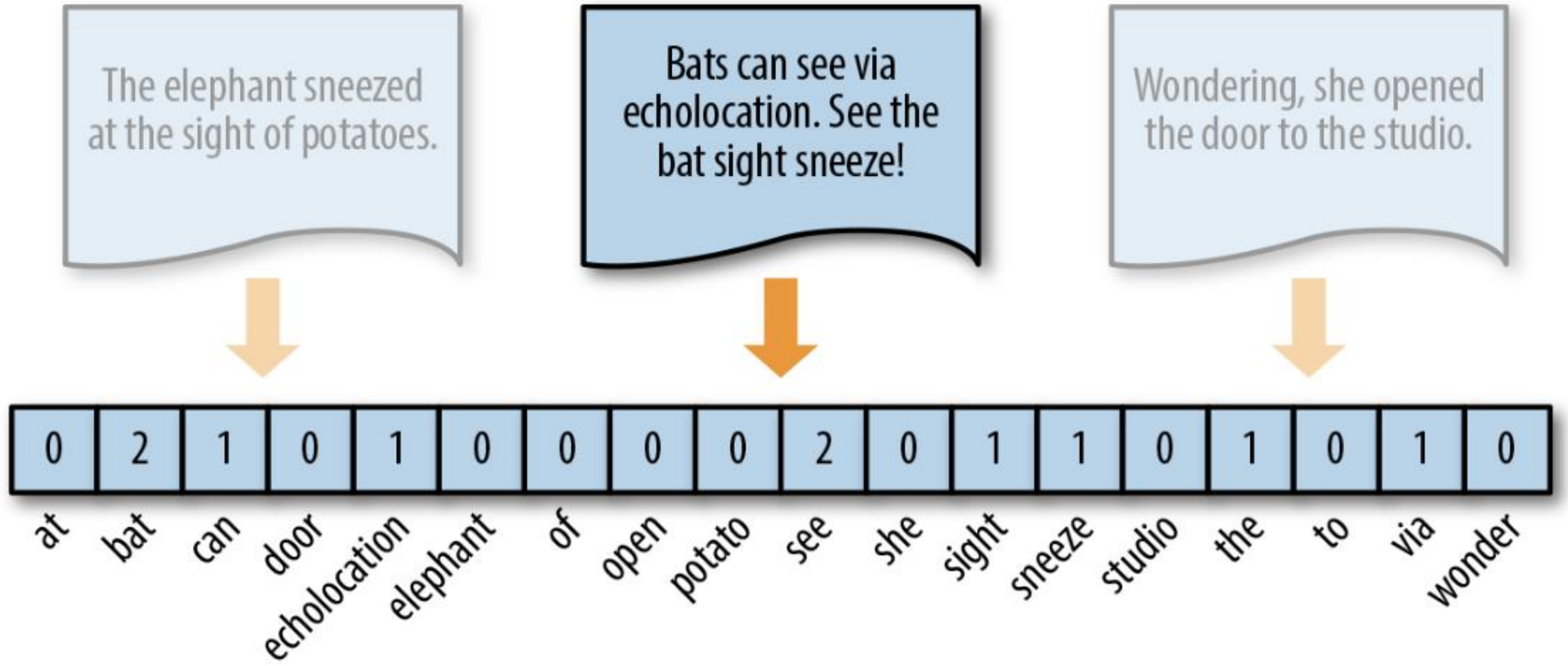


867 muestras

Tecnológico de Monterrey

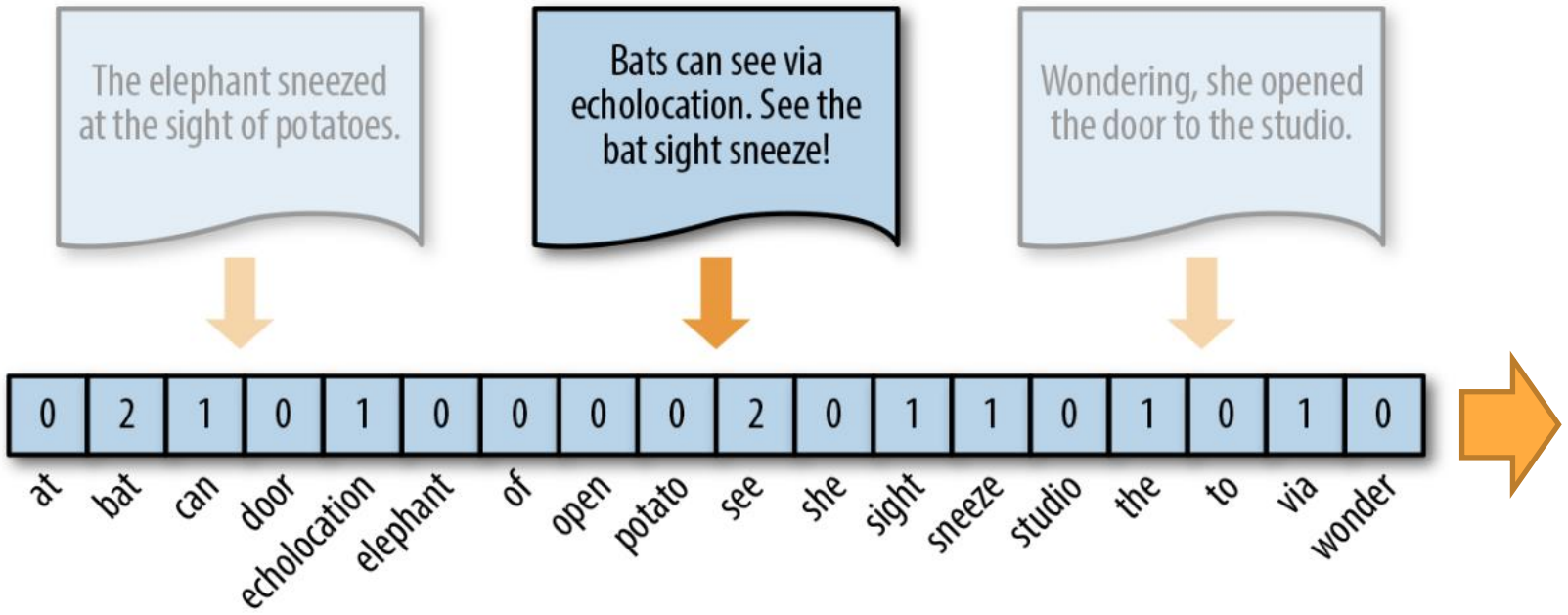
Formato de los vectores

Para esta actividad, se trabajara con un vector de bolsa de palabras (Bag of Words) en el que, en lugar de usar 0 y 1, se utiliza la frecuencia de aparición de las palabras. La siguiente imagen ilustra la representación ha emplear:



Formato de los vectores

Para esta actividad, se trabajara con un vector de bolsa de palabras (Bag of Words) en el que, en lugar de usar 0 y 1, se utiliza la frecuencia de aparición de las palabras. La siguiente imagen ilustra la representación ha emplear:



Es importante señalar que, para esta actividad, se seleccionaran las características a utilizar. Por lo tanto, no se debe tomar la imagen de forma literal en cuanto al número de características a usar en el proceso de entrenamiento.

Tecnológico de Monterrey

Formato de la actividad

- Todas las actividades son individuales.
- La actividad se entregará el día de la siguiente sesión de clases con el profesor.
- Todas las tareas deben estar documentadas en detalle. Las tareas deben entregarse o enviarse como un archivo comprimido (nombre: matricula) con la siguiente información:
  - Archivo de Python, notebook de Jupyter o Google Colab.
  - Documento con explicación del problema e imágenes de su ejecución. Pueden simplemente mostrar un notebook (si lo ocuparon) si toda la documentación está en celdas de markdown, debidamente explicada.

Puede usar Github con markdown para documentar el proceso y solo enviar la URL.

Formato de la actividad

- La actividad deben contener lo siguiente:
  - Portada con información del estudiante (5%).
  - Sección de introducción que explica explícitamente el problema abordado (10%).
  - Sección experimental que explica el enfoque utilizado para resolver el problema (30%).
  - Sección de resultados que recoge los hallazgos hechos(40%).
  - Sección de conclusiones que recoge lo aprendido en la actividad (15%).

Tecnológico de Monterrey

Gracias

Tecnológico de Monterrey