Prueba 1

1. **Dos vectores son ortogonales si son linealmente independientes**
    a. Verdadero
2. **En un problema de aprendizaje automático, el test set se usa para validar los resultados del modelo y no puede ser conocida por el modelo durante su entrenamiento**
    a. Verdadero
3. **Fruta es un hiperónimo de naranja**
    a. Verdadero
4. **Marque lo correcto con respecto a PLSA**
    a. Utiliza un método probabilístico en vez de SVD
5. **Marque lo correcto con respecto a clasificación de textos**
    a. Es un proceso supervisado que asigna una o más categorías a documentos
6. **Marque lo correcto con respecto a la representación Bag of Words**
    a. Es una representación basada en un **espacio vectorial**
7. **Un morfema es una unidad de lenguaje que no puede ser sub-dividida**
    a. Verdadero
8. **Seleccione las técnicas que se relacionan con aprendizaje no supervisado**
    a. Topic Modelling
    b. Hierarchical clustering
9. **Marque lo correcto con respecto a Latent Semantic Analysis (LSA)**
    a. Utiliza SVD
    b. Asume que existe una estructura semántica latente que puede ser descubierta
10. **¿Para que se realiza el proceso de smoothing?**
    a. Para asignar probabilidades distintas de 0 a palabras o n-gramas no vistos previamente
11. **¿Cómo se calcula la métrica de similaridad del coseno?**
    a. La proyección de un vector sobre otro
12. **Marque lo correcto con respecto al etiquetado de Part of Speech (POS)**
    a. También conocido como etiquetado gramático
    b. Es el proceso de etiquetar palabras en un cuerpo de texto que corresponden a una parte del lenguaje en particular
13. **Seleccione la diferencia entre un clasificador generativo y uno discriminativo**
    a. Generativo: aprende el joint probability de p(x,y); Discriminativo: aprende la probabilidad condicional de p(y|x)
14. **En un proceso de Information Retrieval, ¿Qué se realiza en la función de rankeo?**
    a. Se realiza un ordenamiento considerando las necesidades de información y la representación de los documentos
15. **Topic coherence es una métrica para evaluar el modelamiento de tópicos**
    a. Verdadero
16. **Marque los puntos negativos de LSA**
    a. Complejidad computacional cuadrática
    b. El número de temas no es conocido
17. **Seleccione lo correcto con respecto al proceso de Tokenización**
    a. Transforma texto en un conjunto de tokens
18. **Seleccione el concepto correcto de Information Retrieval (IR)**
    a. Representación, almacenamiento, y organización en bases de datos o repositorios y su extracción de acuerdo a necesidades de información
19. **Qué suposición realiza la heurística de Term Frequency**
    a. Un término es más importante si ocurre más frecuentemente en el documento.
20. **En Active Learning, las instancias usadas para entrenar son etiquetadas por un "experto", considerando los ejemplos más difíciles. Seleccione una:**
    a. Verdadero

21. **El Text Summarization se usa principalmente para lograr una representación más compacta de los documentos para su posterior uso en otros métodos automáticos.**
    a. Falso
22. **Las stopwords son palabras que no tienen utilidad para el análisis de documentos**
    a. Verdadero
23. **Marque lo correcto con respecto a variables latentes**
    a. Son variables no, o parcialmente, observadas que puede ser inferidas mediante algún proceso
24. **Text Clustering es una técnica supervisada para agrupar términos o documentos.**
    a. Falso
25. **Qué suposición hace la representación de bag of words**
    a. Las palabras son independientes entre sí

Interciclo

26. **POS se considera como parte de un análisis léxico**
    a. Verdadero
27. **Una representación de lógica de primer orden puede ser utilizada para realizar inferencia lógica en textos**
    a. Verdadero
28. **Un modelo discriminativo intenta explicar únicamente la variable objetivo**
    a. Verdadero
29. **Si se utiliza un modelo de Hidden Markov de segundo orden para realizar etiquetado POS ¿Qué implicaciones tiene?**
    a. La etiqueta actual depende de las 2 etiquetas anteriores
30. **A que hace referencia el término "ambigüedad" en palabras**
    a. Una palabra puede tener varios significados
31. **¿Cuál es la relación entre divergencias léxicas y semánticas?**
    a. Ninguna de las opciones anteriores
32. **¿Existe alguna relación entre LSA y PCA? ¿Explique detalladamente su respuesta?**

Su principal relación es su uso, ambos son aplicados para disminuir o remover el ruido dentro de un conjunto de datos, LSA observa la frecuencia del término asociado. Disminuyen la dimensionalidad de los datos.

33. **¿Para qué se utiliza un cuerpo de texto paralelo con el contexto de traducción de textos?**
    a. Para calcular la probabilidad de que un término tenga un cierto significado
    b. Para realizar la operación de "Word-alignment"
34. **Marque lo correcto con respecto a "Word sense disambiguation"**
    a. Es la tarea de identificar el significado de términos en un contexto
35. **Marque lo correcto con respecto al análisis semántico**
    a. Se lo ejecuta posteriormente al análisis sintáctico y su finalidad es realizar inferencias
36. **Seleccione las características que hacen que NLP sea una tarea altamente compleja**
    a. Ambigüedad en el lenguaje (falta una por que tengo 0,5)
    b. Falta de contexto (talvez)
37. **La semántica distribucional usa el contexto en el que una palabra aparece para medir su similaridad**
    a. Verdadero
38. **¿Cuál es la diferencia entre stemming y lemmatization? ¿Se los debe en un orden específico? ¿En qué situaciones es preferible usar uno u otro?**

(2 puntos obtenidos)

Lemmatizacion hace referencia a encontrar la forma del lema o lexema. Mientras indica la normalización de la palabra (ponis -> pony). Steamming reduce las palabras flexionadas o derivadas a su forma raíz (studies -> studi). *Para la representación de los términos de un texto se debe primero realizar un análisis lexico (lemmatizacion) y luego realizar el steaming, siendo este último opcional*. La lematización siempre debe ser usada para el análisis de los términos dentro del texto, mientras que el steaming es preferible usar cuando el texto es en inglés, así también en plurales y/o adverbios.(palabras que no están en su forma base). Stemming tiene contra de que corta las palabras indiscriminadamente, Lemmatization tiene el contra de requerir diccionarios.

39. **Seleccione el objetivo principal de realizar el etiquetado POS**
    a. Determinar la secuencia de etiquetas más probable en un cuerpo de texto
40. **Considerando el siguiente ejemplo: O=dormir, S={dormía, durmiendo, durmió}**
    a. O es el lemma y S los lexemas
41. **¿Para que se realiza el proceso de extracción de relaciones?**
    a. Ninguna de las otras opciones
42. **¿Por qué un diccionario bilingüe no es suficiente para realizar traducción de textos?**
    a. Por que un diccionario no captura las relaciones complejas entre términos ni su orden en el resultado final
43. **Seleccione lo correcto con respecto a la utilidad de realizar etiquetado POS**
    a. Para realizar análisis de sentimientos
    b. Para realizar parseo sintáctico
44. **En un problema de reconocimiento de entidades, ¿cuál es la importancia de usar POS?**
    a. Por que las etiquetas generadas ayudan a eliminar la incertidumbre sobre la probabilidad de que un término sea una entidad
45. **En un problema de etiquetado POS usando HMM, ¿cuál es la variable latente?**
    a. Las etiquetas de los términos
46. **El parseo sintáctico determina la estructura gramática más probable en un cuerpo de texto**
    a. Verdadero
47. **¿Qué es WordNet?**
    a. Una base de datos léxica en varios idiomas
48. **Las stopwords son palabras que no tienen utilidad para el análisis de documentos**
    a. Verdadero
49. **Problema de etiquetado POS vs problema de clasificación tradicional**
    a. El etiquetado de términos asume una dependencia entre el término y la etiqueta, mientras que en clasificación la dependencia se da entre la clase y variable
50. **Topic coherence es una métrica para evaluar el modelamiento de tópicos**
    a. True

Prueba 2

51. **Con base en una representación de "bag of words", es preferible el uso de un modelo Naive Bayes Binomial**
    a. Falso
52. **¿Por qué se dice que la superficie de decisión producida por el teorema de Bayes es óptima?**
    a. Por que minimiza la posibilidad de obtener errores de tipo I y II (falsos positivos y negativos)

53. Seleccione los problemas relacionados con la alta dimensionalidad en problemas de minería de textos

a. Mayor complejidad computacional
b. En alta dimensionalidad, las instancias se comienzan a parecer más unas con otras

54. Un error de tipo I es equivalente a un falso positivo

a. Verdadero

55. En un problema de clasificación, la función aprendida por el modelo recibe como entrada un vector de características y retorna un valor numérico real

a. Falso

56. ¿Qué suposición hace el clasificador Naive Bayes?

a. Que las variables del problema son condicionalmente independientes entre sí

57. **¿Qué características tiene un modelo de clasificación basado en instancias?**

a. Un modelo de clasificación que aprende una superficie de decisión cuando es invocado

58. Marque lo correcto con respecto a modelos generativos y discriminativos

a. Ambos aprenden sus superficies de decisión durante la etapa de entrenamiento
b. Los modelos generativos modelan el "joint probability" de las entradas y salidas, mientras que los discriminativos se basan típicamente en representaciones espaciales

59. Seleccione lo correcto con respecto a las técnicas de Feature Selection

a. Genera un dataset de menor dimensionalidad que el original seleccionando un conjunto de variables que provean mayor calidad de información

60. Cuando se tenga una representación de datos con texto y otras variables exógenas, que métrica de similaridad sería más adecuada

a. Funciones de Kernel

61. Seleccione lo correcto con respecto a las funciones de "mean proximity"

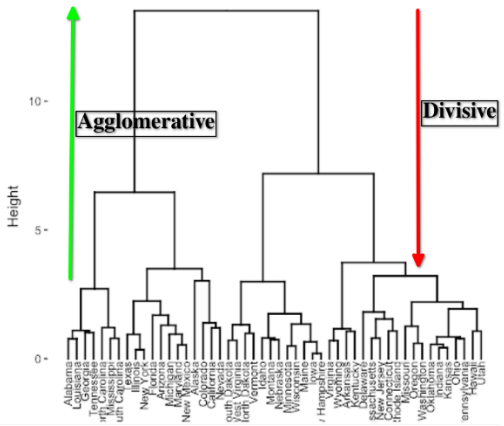
a. Ninguna de las otras opciones

62. **¿Cómo se controlan el número de clusters resultantes en una técnica clustering aglomerativo?**

a. Mediante un threshold con el cual se crean distintos tipos de particiones basándose en el dendograma

63. El algoritmo K-means se lo considera divisivo y aglomerativo al mismo tiempo

a. Falso



64. ¿Cuál es la diferencia entre los algoritmos K-means y K-medoids?

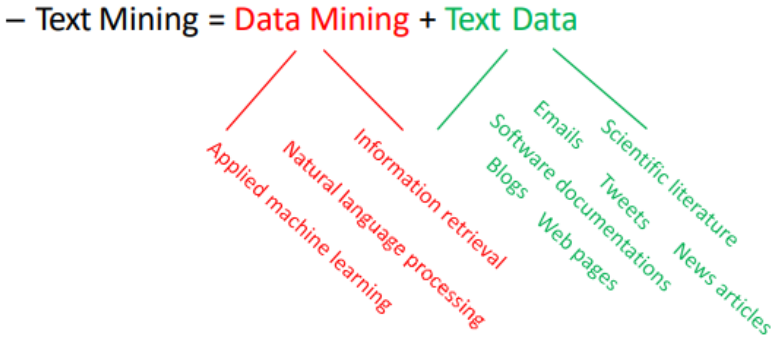
a. Ninguna de las otras opciones

K-means minimiza el error cuadrático total, mientras que k-medoids minimiza la suma de diferencias entre los puntos etiquetados para estar en un grupo y un punto designado como el centro de ese grupo. A diferencia del algoritmo de k-medias, k-medoides elige puntos de datos como centros (medoides o ejemplares).

65. **Seleccione los elementos necesarios para realizar clustering**
    a. Una métrica de distancia/similaridad para comparar elementos
    b. Una representación adecuada de los datos
66. **La función de proximidad máxima compara los objetos más similares entre dos clusters**
    a. Verdadero
67. **Seleccione la técnica correcta que pertenece al tipo de técnica simbólicas**
    a. Reglas de decisión
    b. Arboles de decisión
68. **Con base en una representación de "bag of words", es preferible el uso de un modelo Naive Bayes Binomial**
    a. Falso
69. **¿A qué se refiere un problema de clasificación multi-clase?**
    a. Un problema de clasificación en donde la variable de clase puede tener más de 2 valores
70. **¿Qué es lo que optimiza un modelo de SVM?**
    a. Maximiza la distancia de separación entre los support vectors y la superficie de decisión
71. **¿Por qué se suelen transformar multiplicaciones a sumas con logaritmos en problemas de probabilidad?**
    a. Para evitar problemas de cómputo relacionados con números pequeños
72. **La medida de proximidad "single linkage" es equivalente a una medida de proximidad máxima**
    a. Verdadero
73. **¿Qué características tiene la función objetivo en un algoritmo de clustering?**
    a. Maximiza la distancia entre clusters y minimiza la distancia entre los elementos de cada cluster
74. **Seleccione lo correcto con respecto a las técnicas de Feature Extraction**
    a. Genera un dataset de menor dimensionalidad que el original seleccionando un conjunto de variables que provean mayor calidad de información

LSI = método para la reducción de dimensionalidad usando método de descomposición de valores singulares (SVD).

Cap 1 y 2



Ley de Zipf

- La frecuencia de cualquier palabra es inversamente proporcional a su rango en la tabla de frecuencia.
- Las palabras iniciales toman gran parte de las ocurrencias, pero semánticamente carecen de significado.
- Las palabras finales ocupan una parte importante del vocabulario, pero rara vez aparecen en los documentos.

Normalización

Convertir diferentes formas de una palabra a una forma normalizada en el vocabulario. Ejemplo: U.S.A. -> USA , St. Louis -> Saint Louis

Solución

        - Basado en reglas
            - Eliminar puntos y guiones
            - Todo en minúsculas
        - Basado en diccionario
            - Construir clase equivalente
                - Coche -> "automóvil, vehículo"
                - Teléfono móvil -> "teléfono móvil

- **Stemming/Derivado**
    - • Reducir las palabras flexionadas o derivadas a su forma raíz
        - • Plurales, adverbios, formas de palabras flexionadas
            - • Por ejemplo, damas -> dama
- **Lematización**
    - Encontrar el lema o lexema de una forma de palabra flexionada (el lema es la forma de entrada de la palabra en el diccionario canónico)
- **Porter's algorithm / Porter *Stemmer***
    - Inglés: Tiene 5 fases de reducción de palabras aplicadas secuencialmente
- Dos heurísticas básicas
    - TF (frecuencia de término) = dentro de la frecuencia del documento.
    - IDF (frecuencia inversa de documentos)
- **Ventajas del modelo VS**
    - Empíricamente eficaz
    - Intuitivo
    - Fácil de implementar
    - Bien estudiado / evaluado en su mayoría
- **Desventajas del modelo VS**
    - • Asumir independencia de término
    - • Falta de "adecuación predictiva"
        - • Ponderación arbitraria de términos
        - • Medida de similitud arbitraria
    - • ¡Mucho ajuste de parámetros!

Modelo de Lenguaje: Distribución de probabilidades sobre secuencia de palabras

¿Por qué es útil un LM?

- Proporcionar una forma basada en principios para cuantificar las incertidumbres asociadas con el lenguaje natural.
- Permítanos responder preguntas como:
    - (speech recognition/reconocimiento de voz)
    - (text categorization/categorización de texto)
    - (information retrieval/recuperación de información)
- Mide la fluidez de los documentos

- **¿Por qué solo modelos unigram?**
    - • Dificultad para avanzar hacia modelos más complejos
    - • Implican más parámetros, por lo que se necesitan más datos para estimar
    - • Aumentan significativamente la complejidad computacional, tanto en el tiempo como en el espacio
    - • Es posible que capturar el orden o la estructura de las palabras no agregue tanto valor a la "inferencia temática"

• Pero aún se puede esperar que el uso de modelos más sofisticados mejore el rendimiento

Análisis semántico latente (LSA)

- Los términos / documentos que están estrechamente asociados se colocan uno cerca del otro en este nuevo espacio

- Los términos que no aparecen en un documento aún pueden acercarse a él, si eso es consistente con los principales patrones de asociación en los datos.

LSA probabilístico (PLSA)

- • Método probabilístico en lugar de SVD

Asignación de Dirichlet latente (LDA)

- • LDA = versión bayesiana de pLSA

Capítulo 3: Procesamiento de Lenguaje natural

Morfología: ¿Cuáles son las unidades básicas de significado (palabras)?, ¿Cuál es el *significado de cada palabra*?
Sintaxis: ¿Cómo se relacionan las palabras entre sí?
Semántica:¿Cuál es el "significado combinado" de las palabras?
Pragmática: ¿Qué es el “meta-significado”? (acto de habla)
Discurso: Manejo de una gran cantidad de texto
Inferencia: Darle sentido a todo

Capítulo 4 Modelos de extracción de la información

Definidos por:

- la forma usada para representar el texto del documento y la consulta.
- por el procedimiento de clasificación.

Taxonomía de importantes modelos de recuperación

- Basado en términos
- *Algebraico*
- *Modelo de recuperación Probabilístico*

Capítulo 5 - Text mining en general

Teoría de decisión de Bayes:

- Si conocemos p(y) y p(X|y), la regla de decisión de Bayes es:

$$\hat{y} = argmax_y p(y|X) = argmax_y \frac{p(X|y)p(y)}{p(X)}$$

Riesgo de Bayes:

- **Riesgo:** asignar una clase incorrecta a una instancia
- **Error de tipo I**: falso positivo
- **Error de tipo II:** falso negativo
- Riesgo por la regla de decisión de Bayes puede determinar a una región de rechazo

Enfoques Antiguos: Técnicas Simbólicas

Decision tree/Arbol de Decision:

- Tiene ***nodos y ramas***. Cada nodo, excepto los nodos terminales u hojas, representa una prueba o decisión y se ramifica en subárboles para cada posible resultado de la prueba. Cada hoja esta asociada a una clase particular.

Clasificador Bayesiano: *Calcula la probabilidad posterior de que un objeto nuevo, nunca visto, para que pertenezca a una determinada clase* dadas las características del objeto, esto, *basado en las probabilidades de que estas características individuales estén relacionadas con la clase*.

Función kernel

- Una función de kernel K es un mapeo del espacio de instancia de los ejemplos S a una puntuación de similitud. Una *función del kernel es un producto interno en algún espacio de características*. La función kernel debe ser simétrica, positiva semidefinida y no contiene eigenvalores negativos
- *Funciones típicas del kernel*
  - *lineal (utilizado principalmente en la categorización de texto)*
  - *polinomial*
  - *función de base radial (RBF)*

Ventajas SVM

- SVS puede hacer frente a muchas características (ruidosas): no es necesario seleccionar características a priori, aunque puede seleccionar funciones por razones de eficiencia
- muchos problemas de categorización de texto son linealmente separables

Categorización jerárquica

- Categorías jerárquicas con facetas = un conjunto de jerarquías de categorías, cada una de las cuales corresponde a una faceta diferente y contiene conceptos significativos Ci relevantes para la colección que se va a navegar

Categorización de textos

Pasos generales para la categorización de Texto

1. Construcción y Selección de características
    a. Como representar los documentos de texto
    b. Necesitamos todas las características
2. Especificación del modelo.
3. Estimación y selección de modelos
4. Evaluación

Método de envoltura

- Encuentra el mejor subconjunto de características para un método de clasificación en particular.

Método de Filtrado

Evaluar las características independientemente del clasificador y otras características

Métricas de puntuación de funciones

Frecuencia de Documentos

Palabras raras: no influyen en la predicción global, reducen el tamaño del vocabulario

Ganancia de información

Disminución de la entropía de la reducción categórica cuando una característica está presente frente a ausente

Cross validation

- Evita el ruido en la separación de train/test

Evaluación de clasificación

Accuracy

- Porcentaje de predicción correcta sobre todas las predicciones

Precision

- Fracción de documentos positivos predichos que son de hecho positivos

Recall

- Fracción de documentos positivos que se predice que serán positivos

Precisión : Aplicado a un clasificador para reconocer menos documentos pero con una alta precisión, su valor disminuye a medida que aumenta la cantidad de documentos que se predice que serán positivos.

Recall : Aplicado a un clasificador para conocer más documentos con un mejor valor en sus métricas, su valor aumenta.

Feature selection

Precaución al eliminar funciones de los textos: Las palabras que parecen tener un contenido general bajo pueden ser un indicador de categoría importante, especialmente en combinación con otros términos.

Supervised selection

Calcular la puntuación de relevancia(relevance score) de una característica de texto y seleccionar características con una alta calificación de relevancia. Muchos algoritmos que calculan la correlación entre término y clase, por ejemplo: X^2 (chi-cuadrado), ganancia de información (information gain)

Feature extraction

Transformación de características de texto: por ejemplo,

derivación de palabras (stemming)
formación de frases

Ponderación de términos(weighting) (p. Ej., Frecuencia de términos (tf),frecuencia del documento (idf) y tf x idf

Capitulo 6 Clustering de textos

- Técnica estadística multivariable que permite la generación automática de grupos de datos.
- *Vector de características - feature vector*: usa una matriz multivariable de Xnxp donde:
  - n -> número de objetos a agrupar.
  - p-> número de características medible
- Feature selection y extracción:
  - Elección y ponderación de características: que tan relevante se encuentra una determinada característica con respecto a la agrupación esperada.

Distancia y funciones de similitud:

- Cluster a menudo usan una matriz que indica la distancia o similaridad entre pares de objetos.
- *Funciones simétricas:* Ejem. Distancia euclidiana, coseno…
- *Funciones asimétricas:* Ejem. Kullback-leibler divergence

Funciones de proximidad, entre dos cluster:

- *Proximidad máxima:* basado en pares de objetos más similares.
- *Proximidad mínima:* basado en pares de objetos menos similares.
- *Proximidad promedio (avg):* basado en el promedio de las similitudes entre todos los pares de objetos.
- *Proximidad media (mean):* basado en la similitud del representante (centroide, medioide) de cada cluster.

Algoritmos para cluster:

- **Secuenciales:** En una o algunas iteraciones se construye el cluster.
  - **Algoritmo Single-pass (paso único):** En una pasada todos los n objetos son asignados al cluster más cercano basados en un valor de similitud threshold (umbral).
- **Jerárquicos (Hierarchical)**
  - *Agglomerative clustering:* Comienza desde n objetos individuales, y en los siguientes pasos son agrupados en clusters más generales y finalmente en 1 solo cluster. Específico a lo general.
    - *Métodos difieren en su definición de proximidad entre clusters:*
      - Single link (age)(Vecino más cercano): Usa la función de *proximidad máxima*. Podría generar drawn out clusters
      - Complete link (age)(Vecino más lejano): Usa la función de *proximidad mínima*, tiende a crear clusters muy compactos con diámetro pequeño.
      - Group average link(age): Usa la función de *proximidad promedio*. Genera clusters con forma aprox. de bola.
  - *Divisive clustering***:** Una completa colección de n objetos se divide en grupos cada vez más pequeños hasta que se encuentran los n objetos individuales. General a lo Específico
    - *Hard Clustering:* Para k grupos están: k-means, k-mediod
    - *Soft or fuzzy clustering:* Cada grupo puede pertenecer a diferentes grupos con un grado de pertenencia, cuantificado por coeficiente de pertenencia. Ejm: c-means

Construcción de la matriz de asociación de términos

1. Basado en la coexistencia de términos (o raíces) dentro de los documentos
2. Basado en la co-ocurrencia de términos (o raíces) dentro de los documentos y su distancia (número de palabras entre ellos)

3. Basado en la ocurrencia de términos en vecindarios similares
4. Basado en información puntual mutua (MI) estadístico entre dos términos u y v
5. Based on chi-square value
6. Basado en la log likelihood para una distribución binomial

Otros métodos para el clustering de documentos

- **Agrupación espectral:** el corpus del documento se ve como un grafo no dirigido.
- **Agrupación basada en la factorización matricial no negativa del término por matriz de documento:** *cada eje captura los temas base de un grupo de documentos en particular*, y cada documento se representa como una combinación de los temas base, el tema base más importante determina el grupo al que un documento pertenece

Sistema Scatter/Gather

- Interfaz para navegar por una colección de documentos

Clustering de contenido heterogéneo

- Agrupación de datos de múltiples vistas (minería web, análisis de redes sociales, minería multimedia)
- Agrupación híbrida:
    - Combinar los resultados de múltiples funciones del kernel, por ejemplo, con una interpolación lineal

Cap 7 Resumen del resumen

Sentiment Analysis también se conoce como:

- Minería de opinión, Análisis de los sentimientos, Minería de sentimientos, Detección de subjetividad

Tres métodos para realizar SA:

- Aprendizaje automático: Supervisado / No supervisado
- Basado en léxico: Diccionario / Cuerpo
- Análisis del discurso

Características se pueden expresar como otros problemas de minería de texto

- bolsa de palabras
- n-gramas
- partes del discurso (pe, adjetivos y combinaciones de adjetivo-adverbio)
- palabras de opinión (basadas en el léxico: diccionario o corpus)
- intensificadores y cambiadores de valencia (por negación); verbos modales, dependencia sintáctica

Detección de subjetividad 2 etapas:

- Clasificar como subjetivo o no
- Determina la polaridad

Summarization: Un resumen es un texto que se produce a partir de uno o más textos.
Clasificar como

- *Extractivo*: Los resúmenes extractos se crean reutilizando partes (palabras, oraciones) del texto de entrada literalmente
- *Abstractivo*: La información del texto fuente se reformula, es la forma en la que los humanos realizan resúmenes.

Técnicas de resumen

Supervisado:

- Colección de documentos
- Resúmenes generados por humanos

Sin supervisión o no supervisados:

- El documento es modelado como un gráfico
- Noción de centralidad
- Técnicas no supervisadas:
    - *TextRank:*
        - *Identifica la conexión entre varias entidades, implementa el concepto de recomendación.*

        - *Una unidad de texto recomienda otras unidades de texto relacionadas*
        - *Es más probable que las oraciones muy recomendadas por otras oraciones sean informativas*
    - *LexRank> Las oraciones en el texto se modelan como vértices de un grafo, 2 oraciones están conectadas si existe una similitud entre ellas.*

Visualización de texto

Escenarios típicos de visualizaciones:

- Visualización de una colección de documentos
- Visualización de los resultados de búsqueda
- Visualización de la línea de tiempo de los documentos

Visualizar texto requiere un paso de transformación:

- Discretización
- Agregación
- Normalización

Visualización de documentos de texto TagClouds, por ejemplo, flickr, WordCloud, Treemap, Circle packing

Varias visualizaciones son basado en gráficos

1. Bolsa de palabras
    a. Las palabras en los vectores se ponderan usando TFIDF
2. El algoritmo de agrupación de K-Means divide los documentos en K grupos. Cada grupo consta de documentos similares Los documentos se comparan utilizando la similitud de coseno
3. Los K grupos forman un gráfico:
4. Usando recocido simulado, dibuje un gráfico