Procesamiento de Lenguaje Natural

Topic Modeling

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora

Section 1

Introducción

Topic Modeling

Técnica no supervisada de procesamiento de lenguaje natural que descubre temas ocultos dentro de una colección de documentos, sin necesidad de etiquetado previo.

¿Para qué sirve?

- **Descubrimiento de temas**: Identificar patrones temáticos en grandes colecciones de texto
- **Organización de documentos**: Agrupar documentos por similitud temática
- **Resumen automático**: Extraer los conceptos principales de un corpus
- **Análisis exploratorio**: Entender el contenido de textos no estructurados

Clasificación de técnicas de Topic Modeling

- **Métodos Geométricos**: Basados en la geometría de las representaciones vectoriales y distancias entre features.

  Extracción Features + Clustering

- **Métodos Probabilistas**: Basado en distribuciones de temas y palabras.

  Latent Dirichlet Allocation (LDA)

- **Métodos Algebraicos**: Basado en álgebra lineal y descomposición de matrices de frecuencias o representaciones vectoriales.

  Latent Semantic Analysis (LSA)
  Non-Negative Matrix Factorization (NMF)

Evaluación

La coherencia mide la distancia relativa entre palabras dentro de un tópico. Hay dos tipos principales:

- `c_v`
- `u_mass`

La coherencia de un tópico se define como la suma de puntuaciones de similitud por pares sobre el conjunto de palabras V del tópico:

$$\text{coh}(V) = \sum_{v_i, v_j \in V} \text{score}(v_i, v_j, \epsilon)$$

Section 2

LSA

Latent Semantic Analysis

LSA (Latent Semantic Analysis)

Técnica de procesamiento de lenguaje natural usada en Topic Modelling para descubrir temas en textos, es decir, identificar temas ocultos en un conjunto de documentos.

Fundamentos teóricos

- **Matriz Término-Documento**: Representación numérica de textos, tipicamente BOW o TF-IDF.
- **SVD**: Reducción de dimensionalidad para capturar relaciones semánticas.
- **Espacio semántico latente**: Representación compacta de palabras y documentos.

SVD (Descomposición en Valores Singulares)

- **Propósito**: Reducir la dimensionalidad conservando la estructura semántica.
- **Ecuación**:

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$$

- **Proceso**: Se calculan las matrices simétricas AA^T y A^TA.
- **Componentes**: U son vectores propios de AA^T, V son vectores propios de A^TA, Σ contiene $\sqrt{\lambda_i}$.
- **Reducción**: Se conservan solo los k valores singulares más grandes: $A \approx U_k\Sigma_k V_k^T$.

Espacio semántico latente

- **Concepto**: Representación de palabras y documentos en un espacio de menor dimensión.
- **Ventaja**: Captura relaciones semánticas entre términos y documentos.
- **Ejemplo**: Palabras como "coche" y "automóvil" estarán cerca.

Proceso de LSA

- **Preprocesamiento**: Tokenización, eliminación de stopwords, etc.
- **Matriz Término-Documento**: Creación y ponderación (TF-IDF).
- **SVD**: Aplicación y reducción de dimensionalidad.
- **Interpretación**: Identificación de temas latentes.

Ventajas de LSA

- Captura relaciones semánticas entre palabras.
- Reduce el ruido en grandes conjuntos de datos.
- Simple y fácil de implementar.

Limitaciones de LSA

- Dificultad para interpretar temas explícitamente.
- Depende del preprocesamiento y parámetros.

Aplicaciones de LSA

- Recuperación de información.
- Clasificación de textos.
- Análisis de sentimientos.
- Recomendación de contenido.

Section 3

LDA

Descubrimiento de Temas en Documentos

Problema

¿Cómo organizar documentos automáticamente por temas cuando solo tenemos el texto?

Reto principal

Los algoritmos no entienden el significado de las palabras, solo ven patrones estadísticos

Solución

Latent Dirichlet Allocation (LDA) descubre estructura temática mediante análisis probabilístico

La idea central detrás de LDA

Primer principio

Cada documento es una mezcla de temas

Segundo principio

Cada tema es una distribución de palabras

Metáfora visual

Imagina un triángulo donde cada esquina es un tema puro y los documentos son puntos dentro

¿Cómo "crea" documentos LDA?

Paso 1

Para cada documento: muestrear una distribución de temas

Paso 2

Para cada palabra en el documento:

- Escoger un tema según la distribución del documento
- Escoger una palabra según la distribución del tema

Resultado

Documentos artificiales que imitan la estructura del corpus real

La geometría de los temas

Triángulo temático

Cada esquina representa un tema puro (ciencia, política, deportes)

Posición de documentos

La ubicación dentro del triángulo indica la mezcla temática

Ejemplo

Un documento cerca de la esquina "deportes" trata principalmente de ese tema

Los criterios de optimalidad

Coherencia documental

Los documentos deben ser "monocromáticos" (pocos temas predominantes)

Coherencia léxica

Las palabras deben ser "monocromáticas" (pertenecer a pocos temas)

Balance

Encontrar la coloración que mejor satisfaga ambos criterios simultáneamente

Gibbs sampling: Organizando la habitación

Analogía

Como ordenar una habitación desorganizada objeto por objeto

Mecanismo

Para cada palabra, reasignar su tema considerando:

- Los temas de otras palabras en el mismo documento
- Las asignaciones de la misma palabra en otros documentos

El proceso iterativo

Inicialización

Asignar temas aleatorios a todas las palabras

Iteración principal

Para cada palabra en el corpus:

- Temporalmente remover su asignación de tema
- Calcular probabilidades basadas en el documento y la palabra
- Reasignar tema según estas probabilidades

Convergencia

Repetir hasta que las asignaciones se estabilicen

Evitando problemas con ceros

Problema

Conteos cero impedirían explorar nuevas asignaciones

Solución

Parámetros de suavización α y β

- α: controla la mezcla de temas en documentos
- β: controla la distribución de palabras en temas

Función

Permiten pequeñas probabilidades incluso para combinaciones no observadas

Del algoritmo a la interpretación

Salida del algoritmo

Distribuciones de probabilidad:

- Temas por documento
- Palabras por tema

Rol humano

Interpretar semánticamente los grupos descubiertos

Ejemplo

Si un tema tiene: "balón, gol, equipo, partido", probablemente "deportes"

¿Para qué sirve LDA?

- **Organización documental**: Agrupar automáticamente artículos, noticias o papers
- **Descubrimiento de tendencias**: Identificar temas emergentes en redes sociales
- **Ingeniería de features**: Crear representaciones temáticas para otros algoritmos de ML
- **Sistemas de recomendación**: Sugerir documentos similares basados en similitud temática