INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

DISEÑO E IMPLEMENTACIÓN DE UN ALGORITMO PARA LA DETECCIÓN DE TEMAS PRINCIPALES EN UN CONJUNTO DE DOCUMENTOS CORTOS

OBJETIVO:

- Diseñar e implementar un algoritmo para la detección de temas principales en un conjunto de documentos cortos

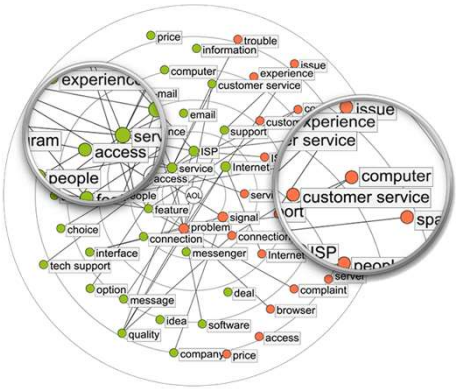

Ilustración con dos temas detectados

RESULTADOS:

- Varias métricas como *purity*, *precision* y *recall* fueron al menos 15% mejor que *LDA* (*Latent Dirichlet Allocation*), un método ampliamente utilizado
- Tiempo de ejecución más favorable que en *LDA*
- Mejor desempeño que el algoritmo *Insight* de la Universidad de Dublín

CONCLUSIONES:

- Se ha construido una herramienta útil
- Tiene sus limitantes (pocos temas, un idioma)
- El análisis de texto ha avanzado mucho, pero resta mucho por hacer

DISEÑO DE LA SOLUCIÓN:

- Limpieza de texto
  - Eliminar *stop-words*
  - Eliminar puntuación
  - Eliminar acentos
  - Corregir o eliminar palabras con faltas de ortografía
- Conversión a vector
- Agrupación jerárquica
- Selección de documentos representativos

IMPLEMENTACIÓN:

- *Tweets* de 140 caracteres o menos
- Lectura y limpieza de *tweets*
- Vectorización, normalización y matriz de distancias
- Doble agrupamiento con *Fastcluster*
- Agrupamiento de *tweets*
- Agrupamiento de grupos de *tweets*
- Relevancia de agrupamientos

Algoritmo final
nuevo educativo modelo equipo
dulce pan cruz azul jesus
comer pan comiendo
muere modelo brasileña criminal case
hambre hoy mañana pan lista
vino pan debe modelo cambiar
veracruz oaxaca pan prd linares
veracruz oaxaca pan stephane suma
avanzar popular permitira nuevo modelo
gustado video @youtube gusto corona

Detección de temas en una colección de tweets con las palabras “pan”, “corona” y “modelo”

REFERENCIAS:

1. Jurafsky, D. and J. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall series in artificial intelligence. Pearson Prentice Hall.
2. Manning, C. and P. Raghavan (2008). *An Introduction to Information Retrieval*. Cambridge University Press

Alumno: Juan Pablo Fonseca Correa
Asesor: Dr. Juan Carlos Martínez Ovando
Carrera: Ingeniería en Computación
Semestre: Primavera 2018