El objetivo principal radica en tomar un número de artículos de economía de distintos países y observar si existe alguna relación entre los mismos, ya que comparten algún patrón regional, además de poder en un futuro clasificar estos diarios de acuerdo al contenido que poseen. Por el tamaño del contenido contemplamos utilizar Mallet ya que permite gestionar un número de artículos altos (más de 10.000), y un poco más rápido que el paquete de textMining de RapidMiner. Para la realización del trabajo práctico se utilizó solr, nutch, scrapy, como software para obtener 500 artículos por diario, en total se obtuvieron 11795 para su entrenamiento y posteriormente un 10% para su prueba y evaluación con un total de 1117. Luego de obtener la data se eliminaron caracteres especiales, como acentos, y se eliminó el contenido del header y footer de los artículos extraídos mediante nutch y solr. Para los artículos extraídos mediante scrapy se eliminaron etiquetas de HTML y caracteres especiales.

1 Introducción

El Modelado de Tópicos es una técnica para tratar documentos que no tienen alguna categorización, esta técnica asume que cada documento es una mezcla aleatoria de categorías o tópicos, donde cada categoría es definida por la preferencia de algunas palabras sobre otras.

Un tópico en el contexto de modelado de tópicos es una distribución de probabilidades de palabras para un conjunto, e indica la probabilidad que una palabra aparezca en un documento sobre un tópico en particular. Para obtener esto el modelado de tópicos asume que las palabras que comprende el texto fueron generadas aleatoriamente y no poseen relación alguna. Su objetivo es Inferir una convincente respuesta bajo la asunción anterior.

Los proyectos de Modelados de tópicos se dividen en:

- *Proyectos de enfoque sincrónico*. en el cual el valor de la unidad de análisis no posee límite de tiempo, o mejor dicho no se identifica con una brecha de tiempo.
- *Proyectos de enfoque diacrónico*. La unidad de analisis de tiempo se genera en un fecha o rango de fecha definido.

El Modelado de Tópicos puede ser usado para clasificar documentos similares además permite mejorar la indexación de texto, pudiendose combinar con métodos de recuperación de información otra de sus aplicaciones consiste en identificar la evolución de ciertos tópicos sobre un periodo de tiempo además de encontrar relaciones entre los diferentes tópicos. También predecir citaciones en base a la presencia de tópicos similares en un texto.

Para evaluar que modelo de tópicos se ajusta mejor a los datos, muchos papers mencionan que es necesario ejecutar en N iteraciones dependiendo del objetivo del modelo si es algo muy general o muy específico depende de los parámetros que reciba el algoritmo. Otra medida es la perplejidad por palabra que consiste en la cantidad de bits utilizados para modelar un texto, en cuanto la perplejidad es menor se podría decir que el modelo es más específico y se ajusta mejor al texto, por lo que se considera mejor tener un modelo con menor perplejidad.

2 Preparación e Instalación

En cuanto a nutch y solr sólo requieren ser descargados y modificar los archivos de configuración a medida de lo que se requiera, en nuestro caso necesitabamos que indexara páginas que pertenecieran al mismo dominio. Además Requiere tener java instalado en la máquina.
El proceso de Descarga de los diarios se realizó durante 2 semanas, sin embargo realizaremos modelado de tópicos sin tomar en cuenta la secuencia en que fueron publicados los artículos. Luego del formato CSV obtenido a través del servidor Solr fue llevado al formato:
Directorio/file1.txt
Directorio/file2.txt

En el cual se importa la carpeta que contiene los archivos que servirán como entrenamiento para el modelado de tópicos. Luego se tomó un 10% como muestra para realizar la evaluación e inferencia de nuevos tópicos. Los archivos usados para entrenamiento se encuentran en la carpeta *train*, y los usados para evaluar el modelo en la carpeta *test* en el repositorio listado en las referencias.

En este trabajo utilizamos Mallet y la implementación que contiene del Algorithm LDA (latent Dirichlet allocation) para realizar el Modelado de tópicos, mediante el ajuste de las frecuencias de palabras a la probabilidad a priori, y Gibbs Sampling para simular el cálculo de la probabilidad a posteriori, sin embargo se puede utilizar otro tipo de distribución en ambas etapas, editando parámetros de software Mallet.

se realizó el Modelado de tópicos en muchas iteraciones con diferentes parámetros para visualizar quré combinación de parámetros sería la más óptima. Durante 4 iteraciónes se añadió más palabras al listado de stopwords, cabe destacar que Mallet no trae en su instalación por defecto con un listado de stopwords en español. Sin embargo se pueden aadir una lista de stopwords en función del corpus.

† Script para formatear Git Hub: [http://bit.ly/15VZJBC].

Cada iteración consta de los siguientes procesos:

- *Importar Archivos*, Este proceso convierte los archivos en formato directorio/file1.txt al formato Mallet, para el entrenamiento y evaluación del Modelo.
- *Entrenamiento*, Este proceso a través de los parámetros configurados en Mallet realiza el entrenamiento con el formato que arrojó el paso anterior. (entrenamiento_unigram.sh)
- *Inferencia de nuevos datos*, Teniendo el archivo con formato Mallet de los datos de test, y el archivo del modelo de entrenamiento, obtenido en el paso anterior, se realiza la asignación de tópicos a los documentos de *test* (ver inferencer.sh)
- *Evaluacion de la inferencia*, Mallet permite evaluar los modelos generados, a través del comando evaluate-topics, esto toma como parámetro de entrada el archivo obtenido en el anterior paso. Y permite visualizar que tan específico o general es un modelo a través de la métrica de perplejidad. (ver evaluate.sh)

Luego de haber cumplido el anterior proceso, graficamos la perplejidad y elegimos dos modelos, k=8 (k = número de tópicos) porque tiene el mayor peso en uno de sus tópicos con respecto a todas las iteraciones y conserva la generalidad del modelo para asociar diarios. Y k=4 e i=10 ya que de acuerdo a la siguiente gráfica tiene el menor valor de la perplejidad en este número de tópico que es más general.

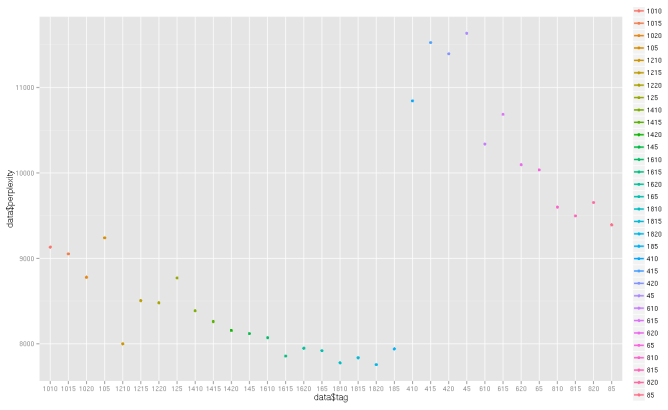

Fig. 1 Perplejidad para cada iteración con la variación de k e i (parámetro que optimiza el modelado de tópico)

Además se puede observar que a medida que incrementamos los tags se decrementa la perplejidad, por lo que es una medida que te permite definir si será mejor clasificador el modelo. Para modelos más generales la perplejidad será mayor.

2.1 Visualizando más de 10000 artículos

Los archivos generados por Mallet luego de su entrenamiento son:

- *noticias_keys.txt*, tiene un listado de los tópicos , y a cada tópico le asigna un peso, de acuerdo a su relevancia. hicimos un gráfico de círculos donde el diámetro del círculo corresponde al peso del tópico.
- *noticias_composition.txt*, hicimos un treemap limitando el número de documentos a incluir por tópico, ya que el número de documentos es muy alto.
- word-topic-count-files.txt, con este archivo realizamos el grafo de la distribución de palabras a través de los tópicos. Este grafo se realizó tomando un límite de probabilidades ya que son muchos documentos.
- noticias_inferencer.txt, muestra el tópico asignado a el nuevo documento.
- *noticias_evaluator.txt*, Permite obtener métricas para evaluar el Modelo, entre ellas la perplejidad.

2.2 Gráficos de burbujas para Tópicos para k=8

Fig. 2 Tópicos y sus Pesos.

El tópico más relevante, de acuerdo a su frecuencia de aparición es el tema del cepo cambiario. Puede deberse a incluir mayor cantidad de diarios de Argentina que del resto del mundo. Otro tópico que destaca, menciona la crisis y deficit en europa. Al hablar de petróleo, reformas, chavéz y crisis se asocia a Venezuela. Y el tópico que muestra temas como tópicos guerra, snowden, amenaza, usd, gas se asocia a Rusia.

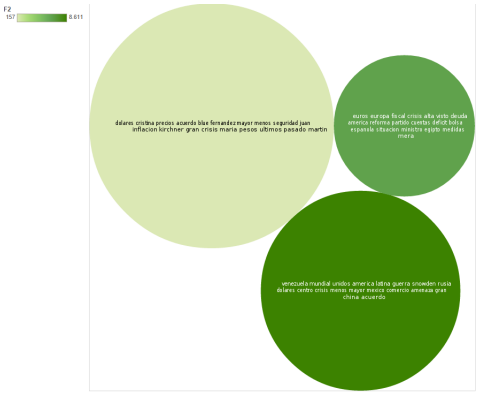

Fig. 3 Tópicos y sus Pesos.

2.3 Gráficos de burbujas para Tópicos para k=4

- Dolares, Blue, fernandez, inflación, crisis, seguridad es el tópico con mayor peso y asocia a todos los diarios de argentina.
- Crisis en europa, deuda y deficit, representan a los diarios españoles.
- Venezuela, mundial, unidos, america latina, guerra, snowden, dolares, crisis, amenaza, acuerdos, china. Asocia a los diarios de actualidadRT.com y avn.info.ve, puede deberse a que ambos países ofrecieron asilo a snowden.

2.4 Grafo con la distribución de palabras por tópicos para k=4

Se puede observar que el tópico 3 contiene un número mayor de palabras con respecto a los otros tópicos. Siento este tópico más general: crisis crecimiento reforma america centro venezuela desarrollo latina ministro plan mayor deuda mundial comercio inversiones economico problemas rol unidos.

2.5 Treemap para k=8

- avn.info.ve, telegrafo.com.ec, eleconomista.com.mx, tienen los tópicos: centro venezuela desarrollo chavez interes america mundial muertos comercio latina politicas acuerdo eeuu redaccion revolucion crisis region snowden unidos.
- ultimahora.com, cincodias.com, los asocia con los tópicos: europa credito asuncion fiscal inversiones rosalia subidas egipto depositos impuestos diaz palacio gibraltar. En este tópico debe asociar a diarios de paraguay sólo que no son listado por no tener una probabilidad de pertenencia por encima de 0.998
- elpais.com, nuevatribuna.com menciona: euros alta crisis america europa fiscal deuda reforma debate barcelona ministro partido medidas deficit paga bruselas francia bankia.
- el resto de los diarios de Argentina los asocia con el tópico 2, dolares precios inflacion cristina blue pesos menos crecimiento productos mayor maria crisis acuerdo oficial demanda bolsa credito argentinos deuda.

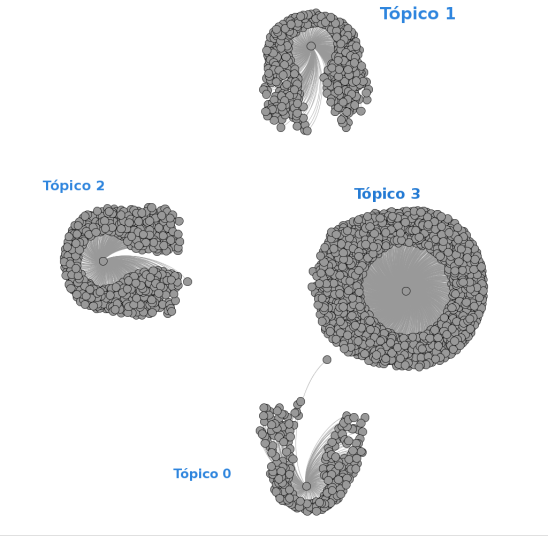

Fig. 4 Distribución de Palabras por tópico.

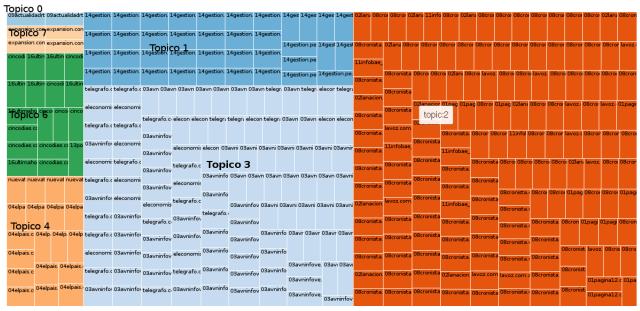

Fig. 5 Treemap para k=8, el color representa un tópico, y como agrupa los diferentes diarios.

3 Conclusiones.

- El potencial del modelado de tópicos no se observa en cada documento individualmente, sino más bien en un

enfoque global analizando grandes cantidades de documentos para visualizar patrones entre ellos.

- El Parametro DirichLet permite darle un peso al tópico, haciendo que sobresalgo por encima de otros tópicos, y la variacion de este parametros permite un mejor ajuste del modelo.
- El intervalo permite que el software realice un calculo optimizado de los parametros alfa y beta cada N iteraciones.
- Conforme se añaden más tópicos el modelo se vuelve más específico, y tiende a colocar todos los documentos que pertenecen a un diario en un mismo tópico, para k=10, ocurrió esto, de 10 tópicos 8 tenía una probabilidad alta por encima de 0.998 y 7 tópicos estaban asociados a un solo diario.
- No existe un modelo mejor que otro, depende del objetivo. Si se desea realizar una clasificador de texto en base al modelado de tópicos, podríamos usar k=cantidad de clases y modificar los parametros que permitan reducir la perplejidad al máximo, de forma que cada tópico pueda representar una clase.
- luego de cada ejecución, siempre arrojará resultados diferentes ya que Mallet utiliza Gibbs sampling por defecto, para calcular la probabilidad a posteriori. Por ende la comparación entre modelos no es muy práctica.
- La cantidad de tópicos recomendada no existe por ende se debe realizar una gran cantidad de iteraciones para observar cual se ajusta mejor a la data. Usualmente mientras más general menor cantidad de tópicos mientras más específico cada tópico tendrá menor cantidad de diarios.

4 Referencias.

- Lista de Desarrollo Mallet http://comments.gmane.org/gmane.comp.ai.mallet.devel/1294
- Modelado de Topicos, Universidad de Princeton http://www.cs.princeton.edu/ blei/topicmodeling.html
- Topic Modelling with MALLET http://hublog.hubmed.org/archives/001870.html
- Comprehending the Digital Humanities https://dhs.stanford.edu/comprehending-the-digital-humanities/
- Gephi + MALLET + EMDA http://www.robincamille.com/
- Identifying the pathways for meaning circulation using text network analysis http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/
- Redes de Tópicos https://dhs.stanford.edu/visualization/topic-networks/
- Review Mallet http://journalofdigitalhumanities.org/2-1/review-mallet-by-ian-milligan-and-shawn-graham/
- Visualizing Topic Models with Force-Directed Graphs http://tedunderwood.com/2012/12/02/visualizing-topic-models-with-force-directed-graphs/
- Mallet Paper http://courses.washington.edu/ling572/winter2013/slides/cl
- SIGKDD 2011 Conferencehttp://www.bytemining.com/2011/08/sigkdd-2011-conference-day-1-graph-mining-and-david-bleitopic-models/
- Código Empleado en este trabajo Práctico http://github.com/j3nnn1/topic_model
- Slides en HTML sobre Modelado de Tópicos http://j3nnn1.github.io/homework/tp_textII/slides/topic_modeling.html
- Perplejidad http://es.wikipedia.org/wiki/Perplejidad