Estudio de la sensibilidad del IPSA con base en la publicación de noticias locales

Proyecto para la Asignatura Captura y Almacenamiento de Datos

Integrantes equipo N°10:

- **Mauricio Bustos**
- **Marco Cid**
- **Oscar Jara**
- **Leonardo Parra**

Problemática / Oportunidad

En una economía abierta como la chilena con una revolución tecnológica - digital cada vez mas pujante, es que la información ha desempeñado un rol importante, especialmente a los actores que la utilizan para invertir en las bolsas de valores debido a su accesibilidad e inmediatez. Este uso de información se ha expresado mediante la volatilidad de precios en instrumentos financieros, especialmente ante circunstancias o sucesos de carácter negativo.

Producto que los movimientos de los mercados financieros responden a diversas variables, tales como: el desempeño de la compañía, el sector al cual pertenecen, las expectativas de futuro y estado anímico de la sociedad, entre otros, es que en su conjunto podrían correlacionarse y generar flujos de inversión o desinversión, explicando alzas o bajas generalizadas en la bolsa local.

Por lo anterior es que dentro de las variables planteadas se observa una oportunidad en el estudio del estado anímico el cual podría ser inferido a partir de la frecuencia o recurrencia de aparición en la prensa local por medio de ciertas palabras claves que generen diversos estados de ánimo.

Dado el contexto, es que en este proyecto se preparará un set de datos que permita estudiar la sensibilidad del IPSA en base a los estados anímicos de la sociedad con el fin de detectar si es posible generar un termómetro de la variable en estudio.

Objetivo General

Preparar la información necesaria que permita estudiar la sensibilidad de precios de cierre del IPSA y el estado anímico de la ciudadanía, mediante la frecuencia de aparición en la prensa local de ciertas palabras claves.

Objetivos Específicos

1. Explorar las fuentes de datos disponibles en la web tanto para los cierres diarios del IPSA como de publicaciones de noticias locales.
2. Extraer y preprocesar ambas fuentes de datos, identificando las necesidades de tratar valores faltantes, outliers, imputaciones y transformación.
3. Consolidar los datasets preprocesados en una sola base integral, en formato JSON, para ser utilizada en estudio posterior.

Fuentes de Datos

1. Valores de cierre históricos del IPSA:
   - Banco Central de Chile: https://www.bcentral.cl/
   - Se obtiene un archivo en formato .xlsx que contiene la fecha y el valor de cierre, según el rango de fechas solicitado.

2. Artículos publicados online de prensa local:
   - News API: https://newsapi.org
   - Se obtiene un archivo JSON con noticias publicadas en web con su fecha, título, descripción, cuerpo, fuente y autor, según palabras claves y el rango de fechas solicitadas.

Etapa 1: Importación de datos

Importación de Noticias ('df_news.xlsx')

- Se definió un conjunto de palabras clave frecuentemente utilizadas en noticias del ámbito económico, y con cada una de ellas se ejecutará una consulta utilizando News API, teniendo como resultado las noticias que incluyen dicha palabra clave publicadas en la web con su fecha, autor, título, contenido y fuente, para un rango de fechas indicada. Algunas de las restricciones de la API son que en cada consulta entrega un máximo de 100 registros, desplegados en páginas de 20 resultados cada una; que el rango de fechas debe ser menor o igual a 30 días; no es posible limitar que los resultados solo pertenezcan a fuentes chilenas.
- Para facilitar el almacenamiento de los resultados, se creó una función personalizada que realiza las consultas y almacena los resultados en un *dataframe*. Se ejecutó esta función personalizada con cada una de las palabras claves y posteriormente se combinaron los resultados en un solo *dataframe*. En cada una de las consultas por palabra clave se añadió la palabra "Chile" para aproximar los resultados a fuentes chilenas.
- El *dataframe* con los registros de noticias fue exportado a un archivo llamado 'df_news.xlsx' para facilitar la revisión posterior sin tener que realizar las consultas nuevamente. Se importa inmediatamente el contenido de dicha planilla a un *dataframe* llamado 'df_news'.
- Se visualizaron cinco registros aleatorios y la dimensión del *dataframe*.

Etapa 1: Importación de datos (cont)

Importación de IPSA ('Canasta_07102021121622.xlsx')

- Desde la página web del Banco Central de Chile, se obtuvo el registro histórico del Índice de Precios Selectivo de Acciones (IPSA) para el periodo comprendido entre 4 de enero de 2021 y el 5 de octubre de 2021. Esta información se descarga en una planilla Excel llamada 'Canasta_07102021121622.xlsx'.
- Se importó esta información a un *dataframe* llamado 'df_ipsa', al que se le eliminaron las primeras tres filas ya que no contienen información, se renombraron las columas, se reinició el índice y se ajustó el formato de la columna fecha.
- Se visualizaron los primeros cinco registros.

Etapa 2: Limpieza de datos

Missing Values en Noticias

- Se buscaron y eliminaron los duplicados en la columna “contenido”, quedando 839 registros únicos.
- Se buscaron los registros que contengan valores perdidos (NaN) en la columna “contenido”, existiendo un registro. Se eliminó y verificó en las dimensiones del *dataframe*, quedando 838 registros.

Missing Values en IPSA

- Se buscó los valores faltantes en ‘df_ipsa’, existiendo 84 datos perdidos, correspondientes a los días de fin de semana y festivos.
- Se imputaron los valores perdidos con el precio del último día hábil de la semana (criterio de negocio).

Outliers en IPSA

- Para descartar la presencia de *outliers*, se realizó verificación visual mediante un gráfico de líneas.

Etapa 3: Transformación

Agrupación de Noticias

- Se ajustó el formato de la columna fecha. Luego se creó una lista para cada fecha en las que se agrupó el contenido de las noticias, y posteriormente se creó el *dataframe* llamado 'df_agrupado' con esta información.

Unión de los sets de datos

- Se combinó 'df_agrupado' y 'df_ipsa' según la fecha en un solo *dataframe* llamado 'df_consolidado', el cual contiene las siguientes columnas:
  - fecha
  - contenido_noticias
  - valor_ipsa

Expresiones regulares para formato, números, puntuación y símbolos

- Se realizó la limpieza de datos por medio de expresiones regulares, de la siguiente forma: transformación de mayúsculas a minúsculas, vocales con tilde a vocales sin tilde, cambio de "ñ" por "nh" y eliminación de números y caracteres especiales. Se creó una nueva columna con la limpieza realizada, manteniendo los datos originales de respaldo.

Etapa 4: Tokenización

- Se realizó tokenización del contenido de las noticias, el cual se almacenó en una nueva columna llamada 'contenido_tokens'.
- Se revisaron las *stopwords* en español y se añadió la palabra "Chile" al set, para evitar la distorsión en los resultados puesto que había sido añadida como palabra clave en todas las consultas.
- Durante el proceso de iteración con la Etapa 6: *Word Representation,* se identificó alta presencia de palabras "chars" y "li", por lo que también fueron agregadas al set de *stopwords.*
- Se procedió a eliminar del contenido de las noticias las *stopwords*, almacenando el resultado en una columna llamada 'contenido_tokens_no_stopwords'.

Etapa 5: *Stemming*

- Se realizó *stemming* en el contenido de las noticias y le almacenó el resultado en una nueva columna llamada 'contenido_stem_tokens'.
- Se inspeccionó el contenido resultante y se decidió no utilizar la variable *stemming* debido a que la transformación genera pérdida de entendimiento de la palabra original.

Etapa 6: Exploración

Word Representation

- Se generó una nube de palabras con la columna 'contenido_tokens_no_stopwords', detectándose la palabra "chars" con alta frecuencia. Se verificó que la causa está en el origen de los datos, cuando al acortar el contenido de las noticias, la API añade al texto la cantidad de caracteres adicionales que no se almacenaron. Se procede a añadir la palabra "chars" como *stopword* y así eliminar esa palabra del contenido.
- Adicionalmente se agregó la palabra "li" como *stopword*, ya que por si sola carece de significado.

Análisis de frecuencia

- Se generó un gráfico con la frecuencia de los 20 Tokens más frecuentes en el contenido de las noticias.

Análisis de longitud de texto

- Se generó un histograma con la Cantidad de Tokens en el eje horizontal y la Cantidad de Días en el eje vertical.

Etapa 7: Exportación

Se generó el archivo 'IPSA_noticias.json' donde las claves son los nombres de las siguientes columnas:

- Contenido original
- Contenido con limpieza a través de expresiones regulares
- Contenido tokenizado
- Contenido sin *stopwords* en español
- Contenido *stemming* con reducción de las palabras a su raíz gramatical
- Fecha
- Valor del IPSA ('valor_ipsa')


UNIVERSIDAD ADOLFO IBÁÑEZ