1 Introducción

2 Objetivos y Metodología

Se propone una investigación enfocada en estudiar y evaluar la aplicación del "Procesamiento de lenguaje natural" o NLP por su sigla en inglés para construir un índice económico.

3 Desarrollo

Lo primero es definir y entender para que sirve un índice económico, desde ahora simplemente "índice". Este se puede definir como un indicador estadístico asociado a un conjunto de instrumentos financieros, en este trabajo sólo se consideran bonos y acciones, matemáticamente se puede definir como una serie de tiempo $(X_t)_{t=1}^N$ indexada por día. También, se denota como índice al conjunto de instrumentos seleccionados para su construcción. El cálculo del índice se puede hacer de variadas formas dependiendo de la naturaleza de los instrumentos que lo componen, generalmente se hace uso de un promedio ponderado de los precios de los instrumentos en el índice. Por ejemplo, suponiendo un conjunto I de instrumentos financieros, claramente finito, podemos computar el valor de nuestro índice al día t como:

$$X_t = \frac{1}{|I|} \sum_{i \in I} P(i)_t.$$

Donde $P(i)$ corresponde a la serie de precios del instrumento i. Este tipo de índices se usan como indicadores de que tan fructífera ha sido la economía, una subida en este valor se puede entender como un alza en los precios y por lo tanto, un "alza del sector económico" y viceversa. Uno de los índices más importantes es el **S&P 500** de Estados Unidos, país que al ser potencia mundial influye en la economía del resto del mundo, este índice se compone de 500 acciones las cuales cotizan en la primera y segunda bolsa de valores más importantes de EE.UU, la de Nueva York o NYSE (New York Stock Exchange) y NASDAQ (National Association of Securities Dealers Automated Quotation), también ubicada en Nueva York, correspondientemente. En el caso de Chile, existe el **IPSA** o Índice de Precio Selectivo de Acciones, este agrupa las 30 acciones con mayor presencia bursátil en el mercado y es reajustado cada año haciendo entrar y salir acciones.

3.1 Replicación IPSA

La primera parte del trabajo realizado se basó en replicar los valores del **IPSA** con los datos disponibles. La información entregada por **LVA** corresponde a los datos de transacciones diarias en la Bolsa de Santiago, agrupados por mes en distintos archivos. Estos datos se ven como se muestra en la Figura 1.

	Fecha	V	C	Cantidad	Nemo	R	Precio	L	MontoPesos	D	Folio	Hora
158255	2020-01-16	35	66	25000	SECURITY	T1	195.00	NaN	4875000	D	220598	14:54:00
187190	2020-01-14	35	35	1492	ENELAM	T1	167.94	OE	250566	D	215503	12:01:39
272243	2020-01-06	66	35	1685	CENCOSUD	T1	1015.10	NaN	1710444	D	214294	12:32:25
47906	2020-01-29	59	85	300	FALABELLA	T1	3170.00	NaN	951000	D	211146	10:47:54
40125	2020-01-29	88	88	320	CCU	T1	7099.90	OE	2271968	D	219069	12:52:06

Figure 1: Muestra de tamaño 5 de datos disponibles para replicación del IPSA, enero 2020.

La columna *Folio* identifica unívocamente a cada transacción, *Precio* indica el precio al cual se transó, *Cantidad* la cantidad de acciones transadas, *MontoPesos* corresponde a la multiplicación de las dos anteriores y *Nemo* es la etiqueta que se le da a la acción en la bolsa, generalmente corresponde

al nombre de la empresa que emite la acción. Algunos de estos Nemos son: BSANTANDER, FALABELLA y ENELCHILE, por ejemplo. El resto de las columnas no será relevante para este estudio.

Remark 3.1. *La mayoría de los datos financieros no incluyen fechas que correspondan a fines de semana o festivos. Esto se debe a que las instituciones que generan estos datos no trabajan en tales días.*

Remark 3.2. *El análisis no estuvo exento de datos dañados. Las primeras aproximaciones al IPSA se caracterizaban por peaks severamente pronunciados o fechas en las cuales los valores obtenidos simplemente no tenían sentido. Estos problemas se deben a que en el caso de no tener el dato, este era reemplazado por un* -1000 *y además, existe un rango de fechas, específicamente en el mes de marzo del* 2020*, en las cuales los datos no fueron llenados correctamente (ver Figura 2). Esta situación fue confirmada por el supervisor y se decide simplemente eliminar estas filas dañadas.*

	Fecha	V	C	Cantidad	Nemo	R	Precio	L	MontoPesos	D	Folio	Hora
425626	2020-03-13	35	35	85.782	ITAUCORP	T1	3055.000	OE	262.064	D	220805	11:36:18
348279	2020-03-17	35	86	9.988	BSANTANDER	T1	30610.000	NaN	305.733	D	225103	12:15:24
309202	2020-03-18	35	58	452.000	CHILE	T1	62690.000	NaN	28.336	D	218817	12:01:57
411736	2020-03-13	70	90	10.102	RIPLEY	T1	275000.000	NaN	-1000.000	D	234759	13:36:25
411639	2020-03-13	70	58	2.511	CMPC	T1	1.515	NaN	-1000.000	D	234856	13:37:40

Figure 2: Muestra de datos dañados.

Lo primero que se necesita es estimar el precio de una acción en un día cualquiera. Para esto nos fijamos en una acción (o Nemo) y fecha cualquiera, a modo de ejemplo tomamos **CCU** y el 15 de abril del 2020. Para este par existen 456 transacciones (ver Figura 3).

	Folio	Precio	MontoPesos	Cantidad	Precio * Cantidad
229	224374	6369.4	1273880.0	200.0	1273880.0
55	231417	6397.0	767640.0	120.0	767640.0
211	225283	6399.0	236763.0	37.0	236763.0
107	228925	6368.4	312052.0	49.0	312051.6
302	221209	6445.0	1031200.0	160.0	1031200.0

Figure 3: Muestra de tamaño 5 de las transacciones asociadas a **CCU** el 15 de abril del 2020.

Se decide calcular el precio del día como un promedio de *Precios* ponderando por *MontoPesos*.

$$X_t = \sum_{i \in I} P(i)_t w(i)$$

donde $w(i)$ es un peso que representa la capitalización de cada instrumento y es invariante con respecto a t.

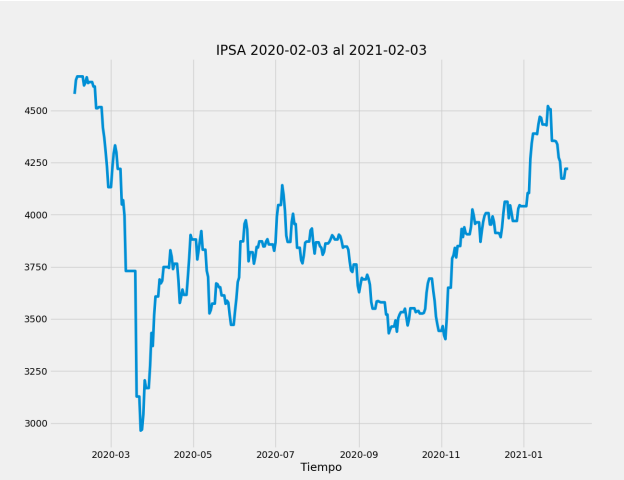

Figure 4: Replicación del ipsa.

4 Punto de que hablar

1. Calculo de índices: IPSA. Explicar el cálculo del índice y por qué se calcula este tipo de indicadores.
2. Análisis de sentimientos con twitter: Hablar de la cantidad de data por nemo y poner gráfico de barras.
3. Hablar de la decisión de dejar retweets porque representan el mismo comentarios pero de otra persona.
4. Comentar que hay tweets en otros idiomas y por lo tanto se hizo necesario trabajar con APIs de idioma.
5. Series de tiempo en economía.
6. Correlación entre dos variables.
7. Test de correlación.
8. Explicar el cálculo del sentimiento saturado, shifted y todo el preproceso que se le hace a los valores pos, neu y neg para transformarlos en un número representativo del sentimiento de cada tweet
9. Explicar que el sentimiento del día asociado a un instrumento se calcula como un promedio de los score de tweets diarios.
10. Explicar que se busca parámetros que maximicen la correlación entre las variables adecuadas: **retorno** del sentimiento, sentimiento promedio en el periodo **sent days**, promedio de los retornos diarios, etc contra el retorno del precio de la acción o menos la tasa de interés para el caso de bonos.
11. Mostrar los gráficos del índice obtenido y los nemos que participan del índice.