Tarea 2: ConvNets (Parte teórica)

Entrega: Jueves 12 de septiembre, 23:59 (Gradescope)

Profesor: Pablo Estévez V.
Auxiliar: Ignacio Reyes J.
Semestre: Primavera 2019

1. Pregunta 2.1

1. Visualice los datos tomando las 2 primeras componentes principales de PCA y compare el resultado de utilizar los 97 tipos productos versus agruparlos en 15 categorías. Compare la varianza explicada por las dos componentes principales en ambos casos.

   Sol:

```
[7]  clustering = KMedians(n_clusters=7)
     pred_labels = clustering.fit_predict(world_data_scaled)
     print("Suma de errores cuadráticos: %f" %(clustering.inertia_))

     Suma de errores cuadráticos: 1945.425474
```

Figura 1: Suma de errores cuadráticos con productos agrupados en 15 categorias

```
[15] clustering = KMedians(n_clusters=7)
     pred_labels = clustering.fit_predict(world_data_scaled)
     print("Suma de errores cuadráticos: %f" %(clustering.inertia_))

     Suma de errores cuadráticos: 17861.859377
```

Figura 2: Suma de errores cuadráticos con productos no agrupados

Analizando los resultados de la suma de errores cuadráticos después de usar el algoritmo K-means, observamos que en el caso de ejemplos no agrupados, hay un error de valor 17861, mientras que en el ejemplo agrupado, el valor cayó sustancialmente, aproximadamente 8 veces, hasta 1945.

```
Varianza explicada por los primeros componentes principales:
[0.15481744 0.10809297]
Suma acumulada de los primeros componentes principales: 0.262910
```

Figura 3: Varianza explicada por los primeros componentes y sú suma acumulada con produtos agrupados

```
Varianza explicada por los primeros componentes principales:
[0.07754024 0.04212973]
Suma acumulada de los primeros componentes principales: 0.119670
```

Figura 4: Varianza explicada por los primeros componentes y sú suma acumulada con produtos no agrupados

Analizando los resultados de la varianza para los dos primeros componentes principales, obtenemos un valor de 0.1196 para el caso no agrupado y 0.2629 para el caso agrupado. Se observó un doble aumento.

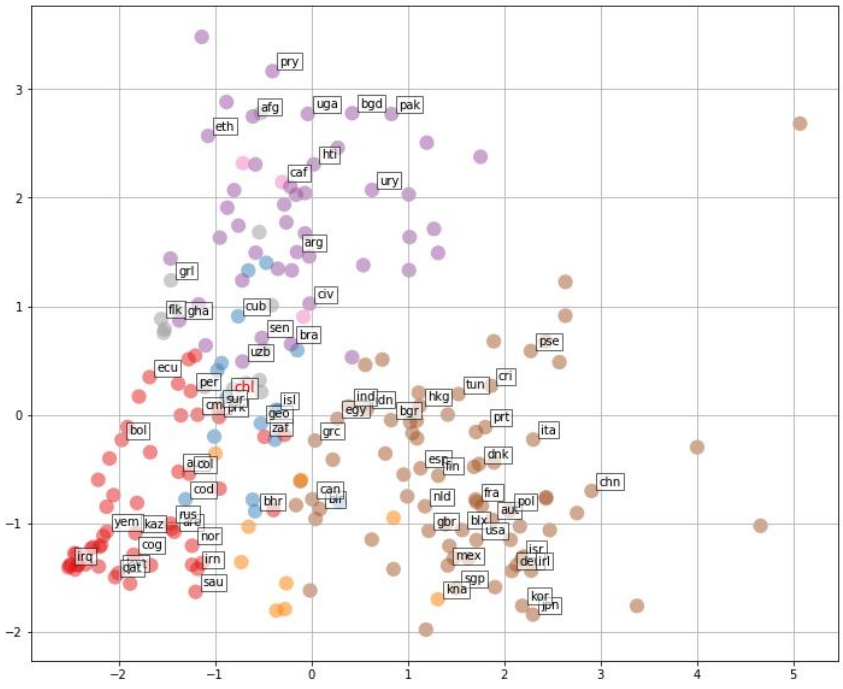

Figura 5: visualización PCA con produtos agrupados

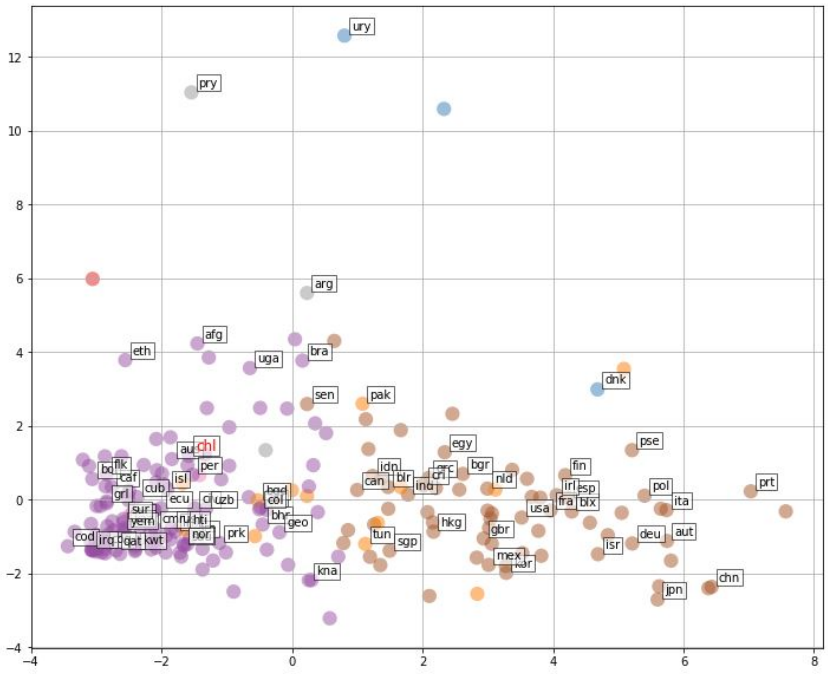

Figura 6: visualización PCA con produtos no agrupados

Al comparar las dos vistas, se puede ver que los grupos son más fáciles de identificar cuando se agrupan. Una conclusión que se puede extraer es que hubo un aumento en la información contenida debido a una mayor varianza.

2. A su juicio, ¿Cuál de las dos visualizaciones revela más información respecto a los datos?

   Sol:

   Como queremos encontrar grupos o similitudes en los datos que tenemos, es más pertinente encontrar las similitudes en los grupos de datos en lugar del conjunto completo. Por lo tanto, cuanto mayor sea la información en los primeros componentes, mejor será la visualización.

   La visualización que revela la mayor cantidad de información es, por lo tanto, la que contiene los productos agrupados en 15 categorías, ya que los primeros 2 componentes principales contienen más variación.

3. ¿Qué dificultades tiene utilizar la información de los 97 productos para describir a cada país?

   Sol:

   El uso de la versión de 97 productos para describir cada país, en lugar de agruparlos, hará que la información o la variación se extienda a más componentes, lo que dificultará la interpretación al reducir el número de componentes principales.

2. Pregunta 2.2

1. Visualice los datos utilizando PCA. Identique y describa los grupos de similitud o clusters entre los países.

   Sol:

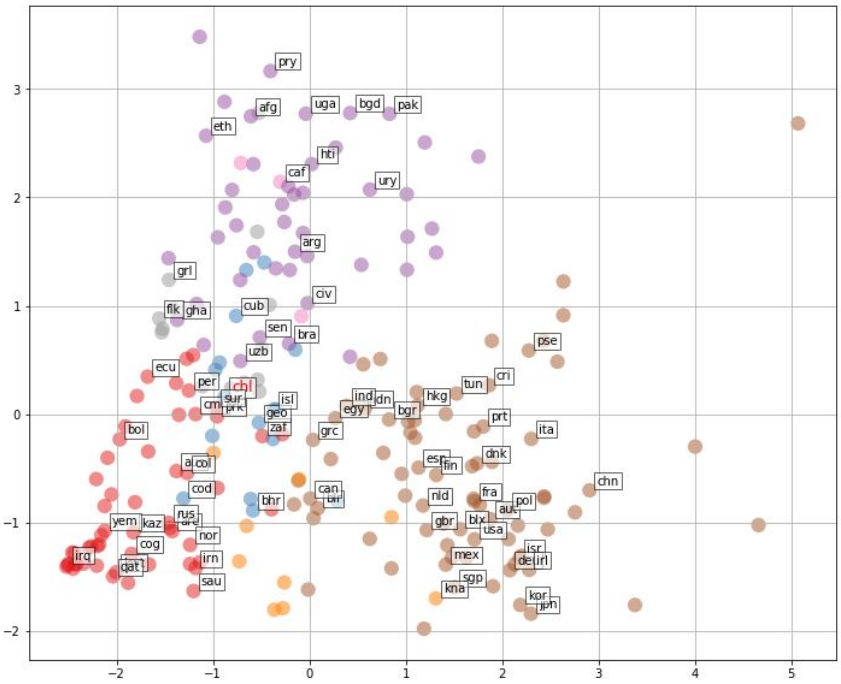

Figura 7: Visualización PCA con produtos agrupados

Color	Países	Tipo de producto que exporta principalmente
Rojo	Iraq, Yemen, Norway	Productos Minerales
Azul	Ethiopia, Afghanistan, Uganda	Productos de origen vegetal
Violeta	Pakistan, Bangladesh,Haiti	Textiles
Marrón	China, Italia, Francia	Máquinas
Rosa	República centroafricana	Madera
Gris	Greenland, Falkland Islands	Productos de origen animal
Naranja	Saint Kitts and Nevis	Equipo Electrico

2. ¿Se observan outliers?

   Sol:

   Un outlier es una observación que está distante del resto de los datos.

   En el grupo de marrones hay 3 o 4 puntos, esto depende de la interpretación, que están sustancialmente distantes del resto del grupo. Por lo tanto, digo que hay 3 o 4 outliers.

3. Pregunta 2.3

1. Compare el resultado de usar PCA versus Kernel PCA con kernel gaussiano. Haga sus comentarios tanto en términos generales como en referencia a clústeres y países específicos.

   Sol:

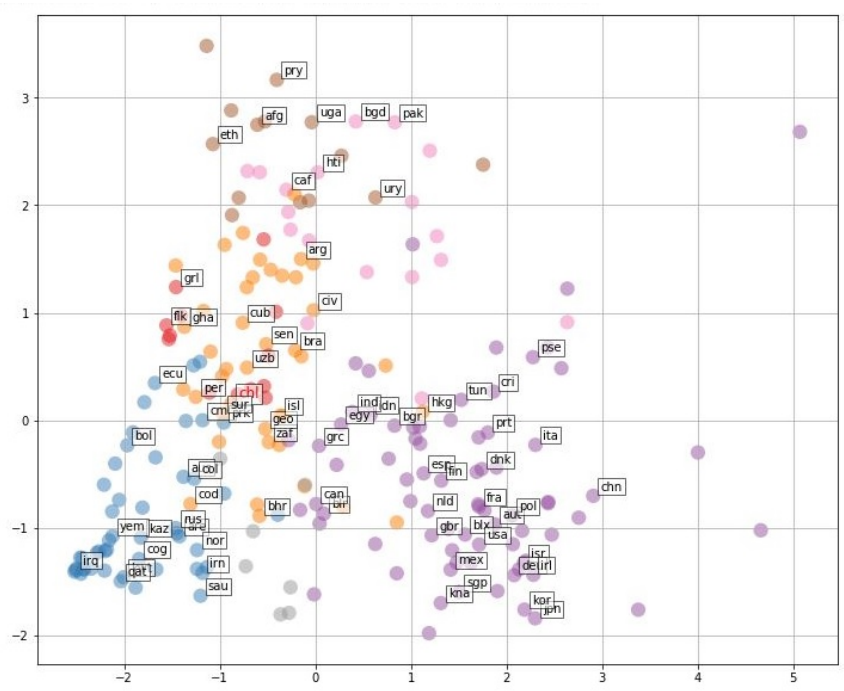

Figura 8: Visualización PCA

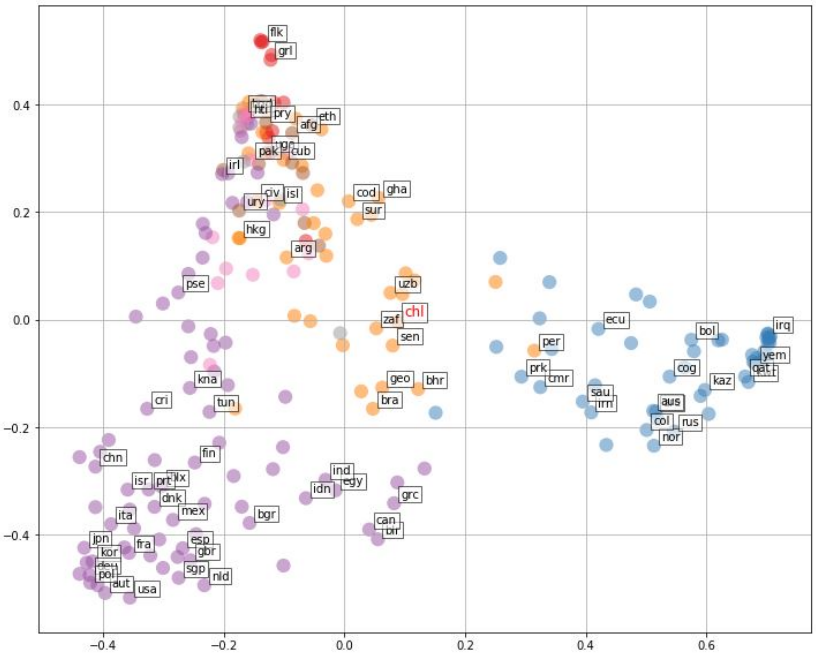

Figura 9: Visualización Kernel PCA

Al comparar las dos vistas, hay diferencias en cómo se organizan, a pesar de mantener el mismo número de grupos.

En el caso del núcleo, se observa que los grupos se agrupan y forman islas, lo que facilita la interpretación y la distinción entre los diferentes grupos.

Países como Yemen e Iraq que estaban en la esquina inferior izquierda en la vista sin kernel se han movido hacia la derecha al usar el método kernel. Italia y Francia que estaban a la izquierda han ido a la derecha y el grupo formado por países como Brasil y Senegal ha ido más alto.

En la parte superior de la vista del núcleo hay varios grupos, como rojo, marrón y rosa, que muestran una similitud entre estos grupos, ya que la distancia entre ellos se correlaciona con su similitud.

4. Pregunta 2.4

1. En PCA sin kernel, elija el número de características que capturen al menos el 85 % de la varianza de los datos. Luego construya un mapa auto-organizativo de Kohonen (SOM) a partir de dicha proyección. Visualice los resultados usando la U-Matrix de distancias. Analice los resultados tal como hizo para los modelos anteriores.

Sol:

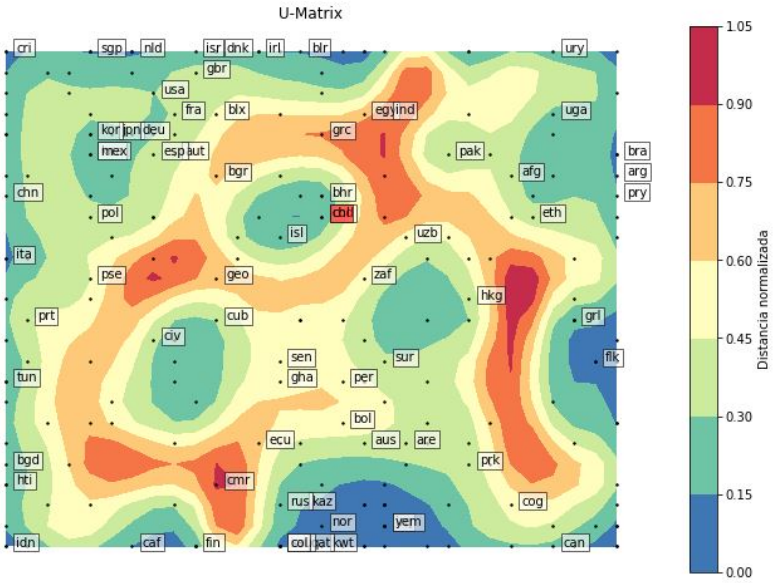

Figura 10: Matriz U correspondiente a SOM aplicado a la proyección PCA

```
Error de cuantización (inicial): 2.630360
Error de cuantización (ajuste grueso): 2.462578
Error de cuantización (ajuste fino): 2.430488
```

Figura 11: Error de cuantización correspondiente a SOM aplicado a la proyección PCA

El número de características para lograr el 85 % de la varianza de datos fue de 11, y acumula una varianza del 88 %.

Para calcular la precisión de la proyección, se utiliza el error promedio de cuantificación en el conjunto completo de datos. Este valor indica qué tan bien se adaptan las neuronas a los datos de Entrenamiento. Como el número de neuronas es menor que el número de datos, el valor del error de cuantificación siempre es mayor que 0.

En estos resultados encontramos que el valor del error de cuantificación es relativamente alto, mayor que 1, con valores de 2.63 (valor inicial) y 2.43 (menor de valores, con ajuste fino).

En U-Matrix, los colores se asignan de acuerdo con el rango de las distancias promedio entre los vectores de peso y sus vecinos. Los colores van del azul al rojo, donde el rojo significa una mayor distancia normalizada, mientras que el azul significa una distancia más corta o una mayor similitud.

En este resultado, es posible afirmar que los bordes están donde hay la distancia más corta, por lo que tendrán más similitudes. También hay 5 grupos con distancias en el rango entre 0.15 y 0.30, en los cuales, por ejemplo, hay un grupo con países como Chile e Islam. También encontramos 4 regiones con o cerca de los colores rojizos, un ejemplo de este grupo es Grecia y Egipto.

5. Pregunta 2.5

1. Usando PCA con kernel gaussiano elija el mismo número de características que usó anteriormente. Luego construya un mapa auto-organizativo de Kohonen (SOM) a partir de dicha proyección. Visualice los resultados usando la U-Matrix de distancias. Analice los resultados tal como hizo para los modelos anteriores.

   Sol:

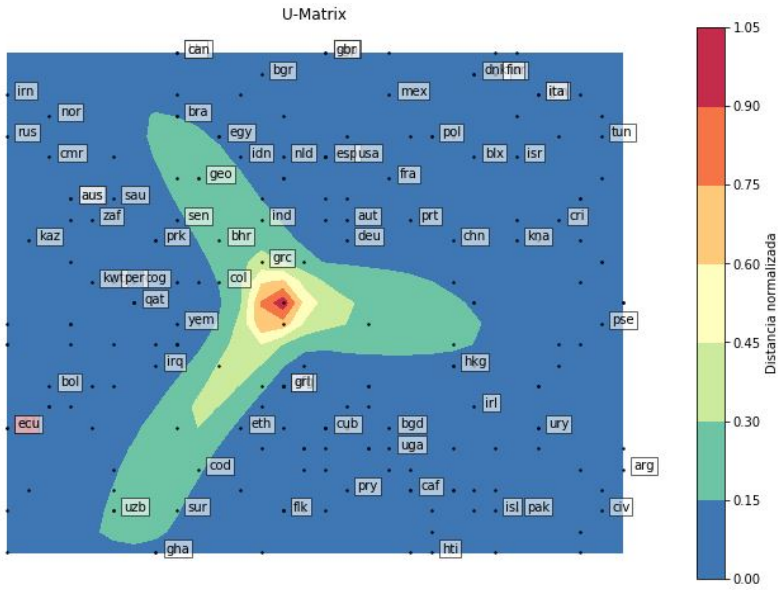

Figura 12: Matriz U correspondiente a SOM aplicado a la proyección Kernel PCA

```
Error de cuantización (inicial): 0.608631
Error de cuantización (ajuste grueso): 0.608624
Error de cuantización (ajuste fino): 0.608621
```

Figura 13: Error de cuantización correspondiente a SOM aplicado a la proyección Kernel PCA

En estos resultados verificamos que el valor del error de cuantificación disminuyó sustancialmente en relación con el caso sin Kernel, presentando un valor de error de cuantificación bajo de 0.61.

El mapa SOM en este caso tiene una distribución de color más centrada radialmente. Las distancias más grandes están en el centro, y a medida que nos alejamos, estas distancias se acortan. En las áreas de color azul, las observaciones experimentan una gran similitud porque se cuida con las más pequeñas distancia entre vecinos.

2. ¿Se observa alguna diferencia al usar la función de kernel?

   Sol:

   Al usar la función con Kernel, notará una gran diferencia en la forma en que se organizan los grupos. En comparación con el resultado anterior, con el uso de los límites del núcleo están mejor definidos y presenta una forma radial donde se pueden identificar fácilmente 3 grandes grupos de alta similitud.

6. Pregunta 2.6

1. Utilizando todas las visualizaciones anteriores indique a qué países se asemeja Chile.

   Sol:

   Al verificar los resultados de todas las evaluaciones interiores, se observa (principalmente en la visualización de clústeres con Kernel PCA y la U-Matrix correspondiente al SOM aplicado a PCA) que Chile permanece cerca de países como Islandia y Bahréin.

   Observando las exportaciones en los datos del Observatorio de Finalización Económica del MIT notamos que estos países son los 3 grandes exportadores de metales.

2. ¿Existe un cluster de países latinoamericanos? ¿Existe un cluster de países árabes / medio oriente?

   Sol:

   No existe un grupo exclusivo de países latinoamericanos o árabes, ya que estos países son mixtos y cercanos a los países de otras regiones que exportan lo mismo.

3. ¿Por qué en algunos casos los clusters tienen coherencia con la ubicación geográfica y en otros no?

   Sol:

   Algunos países tienen coherencia con su ubicación geográfica, porque son países que exportan esencialmente un producto, el petróleo. Este es el caso en países del Medio Oriente como Iraq, Yemen y Kazakstán.

   Sin embargo, en este mismo grupo se encuentran países latinoamericanos como Ecuador y Bolivia. En términos generales, podemos afirmar que este tipo de grupos de países se forman debido a las similitudes en su clima y la abundancia de recursos naturales como los minerales.

   Esto difiere de los grupos formados por países exportadores de equipos eléctricos, por ejemplo, donde se encuentran países de diversas regiones geográficas.

7. Pregunta 2.7

1. ¿A que país se parece México? De acuerdo a su PIB, ¿es un país desarrollado?

   Sol:

   La conclusión a la que llego al observar los datos del Observatorio es que México es un país desarrollado o un país en desarrollo. Su PIB es de $ 18.3k per cápita, similar a los países desarrollados como España y Australia con un PIB de $ 1.31K y $ 1.32K, respectivamente.

2. Compare los tipos de exportaciones de México y Chile, comentando respecto a la diversidad de sus matrices productivas.

   Sol:

   La comparación del número de exportaciones de ambos países muestra que tienen 20 y 21 tipos de exportaciones, respectivamente, Chile y México. Entonces podemos decir que estos países tienen una exportación diversificada. Sin embargo, en México las exportaciones son esencialmente productos de valor agregado, como maquinaria y transporte, que representan el 64 % de las exportaciones, mientras que las exportaciones chilenas se centran más en minerales y metales, que representan el 55 %. de exportaciones.

3. ¿Por qué en algunas visualizaciones aparece un cluster con Groenlandia e Islas Malvinas? ¿Qué tienen en común?

   Sol:

   Dado que ambos son exportadores de productos animales, aparecen juntos en el mismo grupo en algunas de las visualizaciones. Sus porcentajes de exportación son 84 % y 93 % de sus exportaciones, respectivamente Groenlandia e Islas Malvinas.

8. Pregunta 3

1. Visualize la base de datos (agrupando los productos en las 15 categorías) utilizando la técnica t- SNE. Utilize la implementación de T-SNE disponible en sklearn. Analice el resultado obtenido y compare con los experimentos anteriores.

Sol:

```
# ------------------------------------------P3 Programación -- > Visualización con t-SNE
import numpy as np
from sklearn.manifold import TSNE

#perplexity values: 5-50-500

tsne = TSNE(n_components=2,learning_rate=200.0,perplexity=50.0)
tsne_projection = tsne.fit_transform(world_data_scaled)
world_data_projected = tsne_projection

fig = plt.figure(figsize=(12, 10))
ax = fig.add_subplot(1, 1, 1)
ax.scatter(world_data_projected[:, 0], world_data_projected[:, 1],
           c=pred_labels/clustering.n_clusters, linewidth=0, alpha=0.5, s=150, cmap='Set1')
xscale = world_data_projected[:, 0].max() - world_data_projected[:, 0].min()
yscale = world_data_projected[:, 1].max() - world_data_projected[:, 1].min()
for i in range(N):
    if world_labels_short[i] in countries_subset:
        if world_labels_short[i] == "chl":
            ax.annotate(world_labels_short[i],
                        xy=(world_data_projected[i, 0]+0.01*xscale, world_data_projected[i, 1]+0.01*yscale), fontsize=12, color='r',
                       bbox={'facecolor':'white', 'alpha':0.6, 'pad':2})
        else:
            ax.annotate(world_labels_short[i],
                        xy=(world_data_projected[i, 0]+0.01*xscale, world_data_projected[i, 1]+0.01*yscale), fontsize=10,
                       bbox={'facecolor':'white', 'alpha':0.6, 'pad':2})
plt.grid()
```

Figura 14: Codigo para pregunta 3 - T-SNE

Al igual que en las visualizaciones anteriores, se realizó un gráfico de clusters similar al utilizado con PCA, con 2 componentes.

Los resultados obtenidos, con una perplejidad correcta, parecen ser más claros, siendo posible identificar los grupos sin mucha mezcla y sin muchos valores atípicos.

2. Evalúe el impacto de modificar el parámetro perplexity del algoritmo probando un valor pequeño, intermedio y grande.

   Sol:

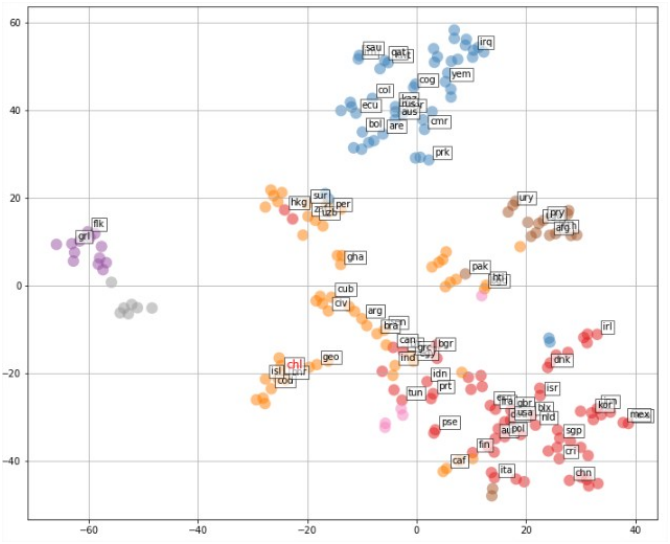

Figura 15: Resultados T-SNE, perplexity=5

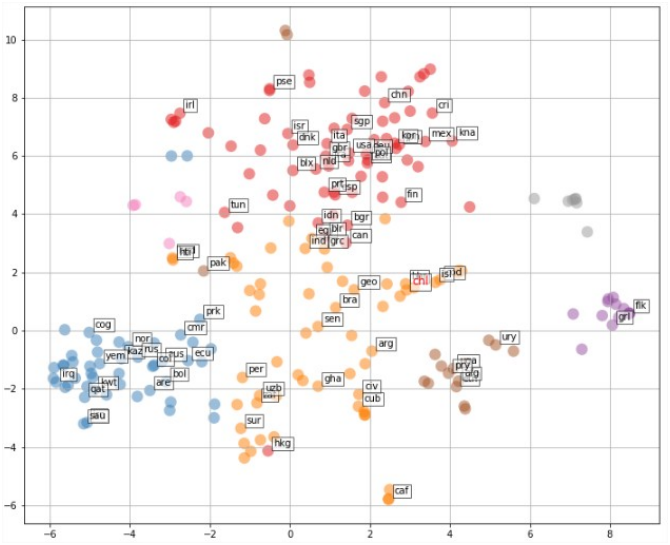

Figura 16: Resultados T-SNE, perplexity=50

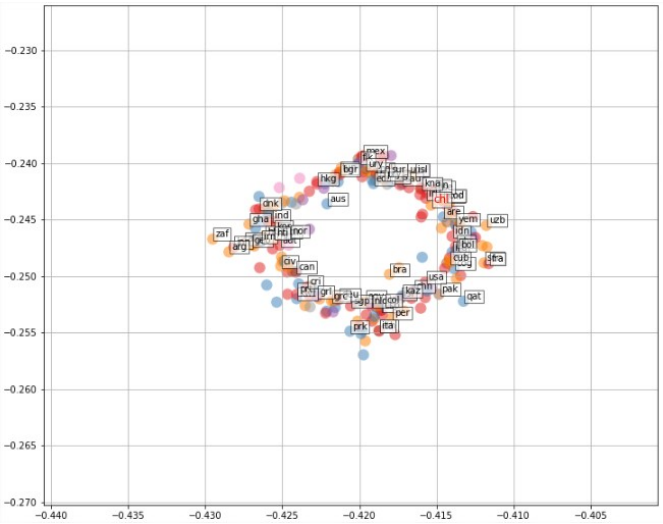

Figura 17: Resultados T-SNE, perplexity=500

Como valores de prueba: pequeño, intermedio y grande, elegimos valores de perplejidad de 5, 50 y 500, respectivamente. La perplejidad es básicamente el número efectivo de vecinos para cualquier punto. Las perplejidades más grandes tendrán en cuenta una estructura más global, mientras que las perplejidades más pequeñas harán que las incrustaciones estén más enfocadas localmente. En este caso, t-SNE funciona relativamente bien para cualquier valor entre 5 y 50.

A diferencia de los valores de alta perplejidad como 500, ya no es posible distinguir los diferentes grupos, presentándolos en forma mixta en forma de anillo.