Part 1:Rule Based NLP and Regex:

```
import re
from nltk.corpus import stopwords
word_to_num = {"one": 1, "two": 2, "three": 3, "four": 4, "five": 5,
"six": 6, "seven": 7, "eight": 8, "nine": 9, "ten": 10}

def generate_bill(text):
    tokens = re.split(r'(?<!\d),(?! \d)|(?<!\d) and', text)
    result = []
    total_bill = 0
    for token in tokens:
        token = ' '.join([str(word_to_num.get(word.strip().lower(),
word.strip())) for word in token.split() if word.lower() not in
['bought', 'kilos', 'each', 'purchased']])
        stop_words = set(stopwords.words('english'))
        token = ' '.join([word for word in token.split() if
word.lower() not in stop_words])
        result.append(token)

    print("Generated Bill:")
    print("{:<20} {:<10} {:<10} {:<10}".format("Product", "Quantity",
"Unit Price", "Total Price"))

    for item in result:
        match = re.match(r'(\d+(?:,\d+)*(?:\.\d+)?) (.+?) (\d+(?:,\
d+)*(?:\.\d+)?)', item)
        if match:
            quantity, product, unit_price = match.groups()
            quantity = float(quantity.replace(',', ''))
            unit_price = float(unit_price.replace(',', ''))
            total_price = quantity * unit_price
            total_bill += total_price
            print("{:<20} {:<10} {:<10} {:<10}".format(product,
quantity, unit_price, total_price))
    print("Total Bill: {:.2f} $".format(total_bill))

text = "I bought three Samsung smartphones 150 $ each, four kilos of
fresh banana for 1,2 dollar a kilogram and one Hamburger with 4,5
dollar"
generate_bill(text)
```

```
Generated Bill:
Product              Quantity   Unit Price Total Price
Samsung smartphones  3.0        150.0      450.0
fresh banana         4.0        12.0       48.0
Hamburger            1.0        45.0       45.0
Total Bill: 543.00 $
```

Part 2: word Embedding :

retrieving the data from mongo DB

```
import pymongo

mongo_client = pymongo.MongoClient('mongodb://127.0.0.1:27017')

db = mongo_client['articles']
db_collection = db.sportArticles

articles =  [article['content']for article in db_collection.find({})]
```

```
['أعلن آياكس أمستردام الهولندي اليوم الثلاثاء، إيقاف أليكس كروس الرئيس التنفيذي الجديد، بشكل فوري، للاشتباه في تورطه بتداولات داخلية لأسهم النادي.وأكّد آياكس في بيان أنّ قرار إيقاف الرئيس التنفيذي جاء بعد أن علم مجلس الإدارة أن كروس اشترى أكثر من 17 ألف سهم من الإعلان عن تعيينه في 2 آب/أغسطس 2023.وكشف النادي أنّ \xa0في آياكس قبل أسبوع الاستشارة القانونية التي أجراها أشارت إلى أنه من المحتمل أن يكون كروس متورطاً في تداول الأسهم من الداخل.وسيتولى أعضاء مجلس الإدارة واجبات ومسؤوليات الرئيس التنفيذي، كما سيقوم مايكل فان براغ (رئيس الاتحاد الهولندي السابق) بمساعدة مجلس الإدارة بشكل موقت.يُذكر أن آياكس يحتل المركز الخامس في ترتيب الدوري الهولندي لكرة القدم برصيد 44 نقطة، بفارق 28 نقطة عن آيندهوفن المُتصدر.\n\n',
 'قدّم روما اليوم الثلاثاء القميص الذي سيرتديه فريق كرة القدم السبت المقبل في مواجهة "ديربي العاصمة" أمام لاتسيو.واستوحي القميص من ذلك الذي ارتداه روما في ديربي موسم 1998-1999 الذي فاز به روما على لاتسيو بثلاثة أهداف لواحد.ويحتضن ملعب "الأولمبيكو" مباراة الديربي السبت المقبل، وهي أهم مباريات الجولة الـ31 من الدوري الإيطالي.\n\n وتهدف مبادرة روما إلى إحياء ذكرى أحد أفضل عصور النادي.وقال النادي في البيان إن القميص القديم "أعيد تصميمه لتكييفه على الأذواق المعاصرة للجيل الجديد من الجماهير".وطرح القميص اليوم الثلاثاء للبيع بسعر 100 يورو و75 يورو للأطفال.',
 'أعلن نادي ريال بيتيس اليوم الثلاثاء، إصابة لاعب وسط الفريق، الأرجنتيني إزيكييل أفيلا في "x" ساقه اليسرى.وأكّد ريال بيتيس في تغريدة في منصة أنّ الفحوصات الطبية التي خضع لها أفيلا أظهرت معاناته من إصابة في العضلات المأبضية بساقه اليسرى.وأصيب أفيلا خلال مباراة فريقه مع جيرونا في الجولة الأخيرة من الدوري الإسباني، والتي انتهت بفوز الفريق "الكتالوني" بنتيجة 3-2.ويحتل ريال بيتيس المركز السابع في ترتيب "الليغا" برصيد 42 نقطة، ويستضيف يوم الجمعة المُقبل، سيلتا فيغو.\n\n',
 'أوضحت المنظمة المصرية لمكافحة المنشطات في بيان، نقاط عدّة بعد إعلان رئيس الاتحاد المصري لكرة القدم جمال علام، ثبوت إيجابية عيّنة لاعب نادي بيراميدز، رمضان صبحي.وجاء في البيان: "لا يجوز أن تصدر بيانات أو تعليقات على أحداث تخص مكافحة المنشطات أو استخدام الشعار الخاص بها إلا من خلال المنظمة، حتى لا تصدر بيانات خاطئة تتسبب في حدوث جدل وتضر بالرياضيين والرياضة المصرية، ومن يخالف ذلك يعرض نفسه للمساءلة".وأضاف "لا تقوم المنظمة المصرية لمكافحة المنشطات بنشر تعليقات أو تصريحات أو نشر الإجراءات التي تقوم بها، وذلك طبقاً للمعيار الدولي للخصوصية والحفاظ على سرية المعلومات".وتابع "من حق المنظمة، وطبقاً للكود الدولي والمعايير الدولية، أن تتواجد في جميع البطولات والمنافسات في جميع الرياضات، سواء كانت فردية أو جماعية لسحب عينات الكشف عن المنشطات. كما يحق لها أن تسحب عينات من أي رياضي في الرياضات الفردية أو الجماعية في أي وقت، وأي مكان خارج المنافسة من دون إخطار مسبق (في المنزل - النادي - المعسكر - التدريب) حسب الخطة الموضوعة من قبل المنظمة المصرية لمكافحة المنشطات طبقاً للمعيار الدولي للاختبارات
```

والتحريات".وكان رئيس الاتحاد المصري لكرة القدم، جمال علام، أعلن أن الاتحاد تلقى خطاباً من المنظمة المصرية لمكافحة المنشطات يفيد بإيجابية عينة للاعب نادي بيراميدز، رمضان صبحي.وفي وقتٍ لاحق، أعلن بيراميدز رغبة رمضان صبحي في مقاضاة جمال علام "بسبب التصريحات غير المسؤولة والخاطئة".',
'كشف نادي الهلال السعودي، أمس الاثنين، طبيعة إصابة مُهاجمه الصربي ألكسندر ميتروفيتش. إنّ الأشعة التي أجراها "x" ومدّة غيابه عن الملاعب.وقال الهلال في تغريدة في منصة ميتروفيتش "أكدت تعرّضه لإصابة في العضلة الخلفية، وسيخضع على إثرها لبرنامج علاجي وتأهيلي لمدة 6 أسابيع".وأصيب ميتروفيتش خلال مباراة فريقه أمام الشباب في الجولة الماضية من دوري روشن السعودي، والتي انتهت بفوز "الزعيم" بنتيجة 4-3.يُذكر أنّ الهلال يتصدّر ترتيب الدوري السعودي برصيد 71 نقطة، بفارق 12 نقطة عن أقرب ملاحقيه، النصر.\n\n',
'فاز الأهلي على ضيفه الاتحاد، بنتيجة 1-0، الاثنين، ضمن منافسات الجولة الـ26 من دوري روشن السعودي لكرة القدم.سجل هدف المباراة الوحيد لنادي الأهلي فراس البريكان في الدقيقة الـ34، لتبقى النتيجة على حالها حتى صفارة النهاية، ويحقّق "الملكي" العلامة الكاملة.ورفع الأهلي رصيده بهذا الانتصار إلى 51 نقطة في المركز الـ3، وبفارق 5 نقاط عن صاحب المركز الـ4 الاتحاد.\n\n',
'فاز أتلتيكو مدريد على مضيفه فياريال، بنتيجة 2-1، اليوم الاثنين، ضمن منافسات الجولة الـ30 من دوري الدرجة الأولى الإسباني "الليغا".أنهى أتلتيكو مدريد شوط المباراة الأول متقدماً على فياريال، بهدف سجله اللاعب البلجيكي أكسيل فيتسل في الدقيقة الـ9.وفي الشوط الثاني، أدرك المهاجم النرويجي ألكسندر سورلوث التعادل لفياريال في الدقيقة الـ51، وسجل ساول نيغويز الهدف الثاني "للروخي بلانكوس" في الدقيقة الـ88.ورفع أتلتيكو مدريد رصيده بهذا الانتصار إلى 58 نقطة في المركز الـ4، وتجمد رصيد فياريال عند النقطة الـ38 في المركز الـ10.\xa0',
'فاز إنتر ميلانو على ضيفه إمبولي، بنتيجة 2-0، اليوم الاثنين، ضمن منافسات الجولة الـ30 من دوري الدرجة الأولى الإيطالي "السيري آ".أنهى إنتر ميلانو الشوط الأول متقدماً على إمبولي، بهدف سجله الإيطالي فيديريكو ديماركو في الدقيقة الـ6.وفي الشوط الثاني، أضاف أليكسيس سانشيز الهدف الثاني" للنيراتزوري" في الدقيقة الـ81، لتبقى النتيجة على حالها حتى صفارة النهاية.ورفع إنتر ميلانو رصيده بهذا الانتصار إلى 79 نقطة في صدارة الترتيب، وتجمد رصيد إمبولي عند النقطة الـ25 في المركز الـ18.\n\n\n\n',
'انتقد رئيس اللجنة الأولمبية الروسية، ستانيسلاف بوزدنياكوف، اليوم الاثنين، تصريحات عمدة باريس، آن هيدالغو، والمتعلقة بمشاركة الرياضيين الروس والبيلاروس في دورة الألعاب الأولمبية، "باريس 2024".وأشارت هيدالغو، في وقت سابق، إلى أنه لن يتم الترحيب بالرياضيين الروس والبيلاروس في دورة الألعاب، في وقت أعربت عن دعمها للرياضيين الأوكرانيين.وقال رئيس اللجنة الأولمبية الروسية إنه في حال لم تكن المدينة المضيفة مستعدة لاستضافة الرياضيين الذين حصلوا على حق المشاركة في المسابقات، فيجب على اللجنة الأولمبية الدولية نقل الألعاب إلى مكان آخر.وأشار بوزدنياكوف إلى أنه ليس من المهم إذا كانت كلمات هيدالغو هذا لا يغير \xa0تندرج ضمن شعار غير مسؤول مناهض لروسيا، أو بيان رسمي، مشيراً إلى أن الجوهر الأساسي، وهو "أنهم حقا لا يريدون رؤية حتى رياضيين حياديين من روسيا في الألعاب الأولمبية".\n\n\n\n',
'تعادل روما مع مضيفه ليتشي، من دون أهداف (0-0)، اليوم الاثنين، ضمن منافسات الجولة الـ30 من دوري الدرجة الأولى الإيطالي، "السيري آ".ورفع روما رصيده بهذا التعادل إلى 52 نقطة في المركز الـ5، وبات رصيد ليتشي 29 نقطة في المركز الـ13.وسيخوض فريق المدرب دانييل دي روسي "ديربي" العاصمة الإيطالية في الجولة المقبلة عندما يستضيف لاتسيو، في حين سيزور نادي ليتشي ملعب "سان سيرو" ليواجه ميلان.\n\n',
'فرض التعادل الإيجابي 1-1 نفسه على مباراة أودينيزي مع مُضيفه ساسولو، اليوم الاثنين، ضمن منافسات الجولة الـ30 من الدوري الإيطالي لكرة القدم، "السيري آ".بادر أصحاب الأرض

```
إلى افتتاح التسجيل عبر غريغوري ديفريل في الدقيقة الـ41، وبعدها بـ3 دقائق سجّل فلوران توفين هدف التعادل لمصلحة أودينيزي، واستمرت النتيجة على حالها لتنتهي المباراة بنتيجة 1-1.وتقاسم الفريقان نقاط المباراة، نقطة لكل منهما، ورفع أودينيزي رصيده إلى 28 نقطة في المركز الـ14، بفارق نقطة واحدة عن هيلاس فيرونا صاحب المركز الـ15، والذي تعادل مع كالياري صاحب المركز الـ16 (27 نقطة) بهدف لمثله.وبقي ساسولو الذي تلقّى خسارته الـ18 هذا الموسم، في المركز الـ19 برصيد 24 نقطة.',
 ' اختير مُهاجم نادي النصر، البرتغالي كريستيانو رونالدو، اليوم الاثنين، أفضل لاعب في دوري روشن السعودي عن شهر آذار/مارس الماضي.تفوّق رونالدو في التصويت على المغربي عبد الرزاق حمد الله مهاجم الاتحاد، والكاميروني كارل إيكامبي لاعب الاتفاق.ويعتلي رونالدو صدارة هدّافي الدوري السعودي هذا الموسم برصيد 26 هدفاً، بفارق 4 أهداف عن أقرب ملاحقيه، الصربي ألكساندر ميتروفيتش مهاجم الهلال.يُشار إلى أنّ رونالدو سجّل ثلاثة أهداف محقّقاً الهاتريك" في انتصار فريقه على الطائي بنتيجة 5-2 في الجولة الماضية".\n\n']
```

tokenizing the data

```
from nltk.tokenize import word_tokenize
tokenized_articles = []

for article in articles:
    tokenized_articles.append(word_tokenize(article))
tokenized_articles[:1]
```

```
[['أعلن',
  'آياكس',
  'أمستردام',
  'الهولندي',
  'اليوم',
  'الثلاثاء،',
  'إيقاف',
  'أليكس',
  'كروس',
  'الرئيس',
  'التنفيذي',
  'الجديد،',
  'بشكل',
  'فوري،',
  'للاشتباه',
  'في',
  'تورطه',
  'بتداولات',
  'داخلية',
  'لأسهم',
  'النادي.وأكّد',
  'آياكس',
  'في',
  'بيان',
```

```
  'أنّ',
  'قرار',
  'إيقاف',
  'الرئيس',
  'التنفيذي',
  'جاء',
  'بعد',
  'أن',
  'علم',
  'مجلس',
  'الإدارة',
  'أن',
  'كروس',
  'اشترى',
  'أكثر',
  'من',
  '17',
  'ألف',
  'سهم',
  'في',
  'آياكس',
  'قبل',
  'أسبوع',
  'من',
  'الإعلان',
  'عن',
  'تعيينه',
  'في',
  '2',
  'آب/أغسطس',
  '2023.وكشف',
  'النادي',
  'أنّ',
  'الاستشارة',
  'القانونية',
  'التي',
  'أجراها',
  'أشارت',
  'إلى',
  'أنه',
  'من',
  'المحتمل',
  'أن',
  'يكون',
  'كروس',
```

```
  'متورطاً',
  'في',
  'تداول',
  'الأسهم',
  'من',
  'الداخل.وسيتولى',
  'أعضاء',
  'مجلس',
  'الإدارة',
  'واجبات',
  'ومسؤوليات',
  'الرئيس',
  'التنفيذي،',
  'كما',
  'سيقوم',
  'مايكل',
  'فان',
  'براغ',
  '(',
  'رئيس',
  'الاتحاد',
  'الهولندي',
  'السابق',
  ')',
  'بمساعدة',
  'مجلس',
  'الإدارة',
  'بشكل',
  'موقت.يُذكر',
  'أن',
  'آياكس',
  'يحتل',
  'المركز',
  'الخامس',
  'في',
  'ترتيب',
  'الدوري',
  'الهولندي',
  'لكرة',
  'القدم',
  'برصيد',
  '44',
  'نقطة،',
  'بفارق',
  '28',
```

```
  'نقطة',
  'عن',
  'آيندهوفن',
  'المُتصدر',
  '.']]
```

removing stopwords and non meaningful charachters

```
from nltk.corpus import stopwords

stopwords = set(stopwords.words("arabic"))
filtred_content = [word for article in tokenized_articles for word in
article if not word in stopwords and not word.isdigit()]
print('before filtration: ',len([word for article in
tokenized_articles for word in article]))
print('after filtration: ',len(filtred_content))
```

```
before filtration:  1215
after filtration:  980
```

```
import re
pattern = re.compile(r'\d')
digit_words = [word.string for w in filtred_content for word in
pattern.finditer(w)]
ls = ['(', '،', ')', '.', '/', ':', '؛',"" ,'ل','ٰ','ب','؟','``',"'`"]
filtred_content2 = [word for word in filtred_content if not word in
digit_words and not word in ls]
print('before filtration: ',len(filtred_content))
print('after filtration: ',len(filtred_content2))
```

```
before filtration:  980
after filtration:  867
```

one hot encoding the data using pandas get_dummies method

```
import pandas as pd
one_hot_encoded_data = pd.get_dummies(filtred_content2)
one_hot_encoded_data
```

```
         -  .وطرح .ورفع .وتابع .وأضاف .وأصيب .وأشارت .بادر .أنهى \
0    False  False  False  False  False  False  False  False
False
1    False  False  False  False  False  False  False  False
False
2    False  False  False  False  False  False  False  False
False
3    False  False  False  False  False  False  False  False
False
4    False  False  False  False  False  False  False  False
False
```

```
..     ...    ...    ...      ...     ...     ...     ...    ...    ..
.
862  False  False  False    False   False   False   False  False
False
863  False  False  False    False   False   False   False  False
False
864  False  False  False    False   False   False   False  False
False
865  False  False  False    False   False   False   False  False
False
866  False  False  False    False   False   False   False  False
False

     \ وكان ... يحتل  يحق يخالف يريدون يستضيف يعرض يغير يفيد.
0    False  ...  False  False  False   False   False  False  False
False
1    False  ...  False  False  False   False   False  False  False
False
2    False  ...  False  False  False   False   False  False  False
False
3    False  ...  False  False  False   False   False  False  False
False
4    False  ...  False  False  False   False   False  False  False
False
..     ...  ...    ...    ...    ...     ...     ...    ...    ...
...
862  False  ...  False  False  False   False   False  False  False
False
863  False  ...  False  False  False   False   False  False  False
False
864  False  ...  False  False  False   False   False  False  False
False
865  False  ...  False  False  False   False   False  False  False
False
866  False  ...  False  False  False   False   False  False  False
False

      يكون  يوم
0    False  False
1    False  False
2    False  False
3    False  False
4    False  False
..     ...    ...
862  False  False
863  False  False
864  False  False
865  False  False
866  False  False
```

```
[867 rows x 541 columns]
```

word2count

```
word2count = {}
for word in filtred_content2:
    if word not in word2count.keys():
        word2count[word] = 1
    else:
        word2count[word] += 1
{k:word2count[k] for k in list(word2count)[:10]}
```

```
{'أعلن': 4,
 'آياكس': 4,
 'أمستردام': 1,
 'الهولندي': 3,
 'اليوم': 10,
 'الثلاثاء،': 2,
 'إيقاف': 2,
 'أليكس': 1,
 'كروس': 3,
 'الرئيس': 3}
```

TF-IDF

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer()
result = tfidf.fit_transform(filtred_content2)
print('\nidf values:')
for ele1, ele2 in zip(tfidf.get_feature_names_out(), tfidf.idf_):
    print(ele1, ':', ele2)
print('\nWord indexes:')
print(tfidf.vocabulary_)
print('\ntf-idf value:')
print(result)
print('\ntf-idf values in matrix form:')
print(result.toarray())
```

Appliying Word2Vec CBOW model using Gensim and FastText

ArWordVec is a set of pre-trained word embedding models derived from Arabic tweets. It aims to support Arabic NLP research. Various techniques were used to build these models, including word2vec and GloVe. Models are named in the format 'model-d-w-m', such as CBOW-500-5-400-10, indicating the approach, vector size, window size, and minimum word count, respectively.

to download the model visit this link

for the fasttext model it was trained on Common Crawl and Wikipedia using fastText. These models were trained using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives. to download the model visit this link

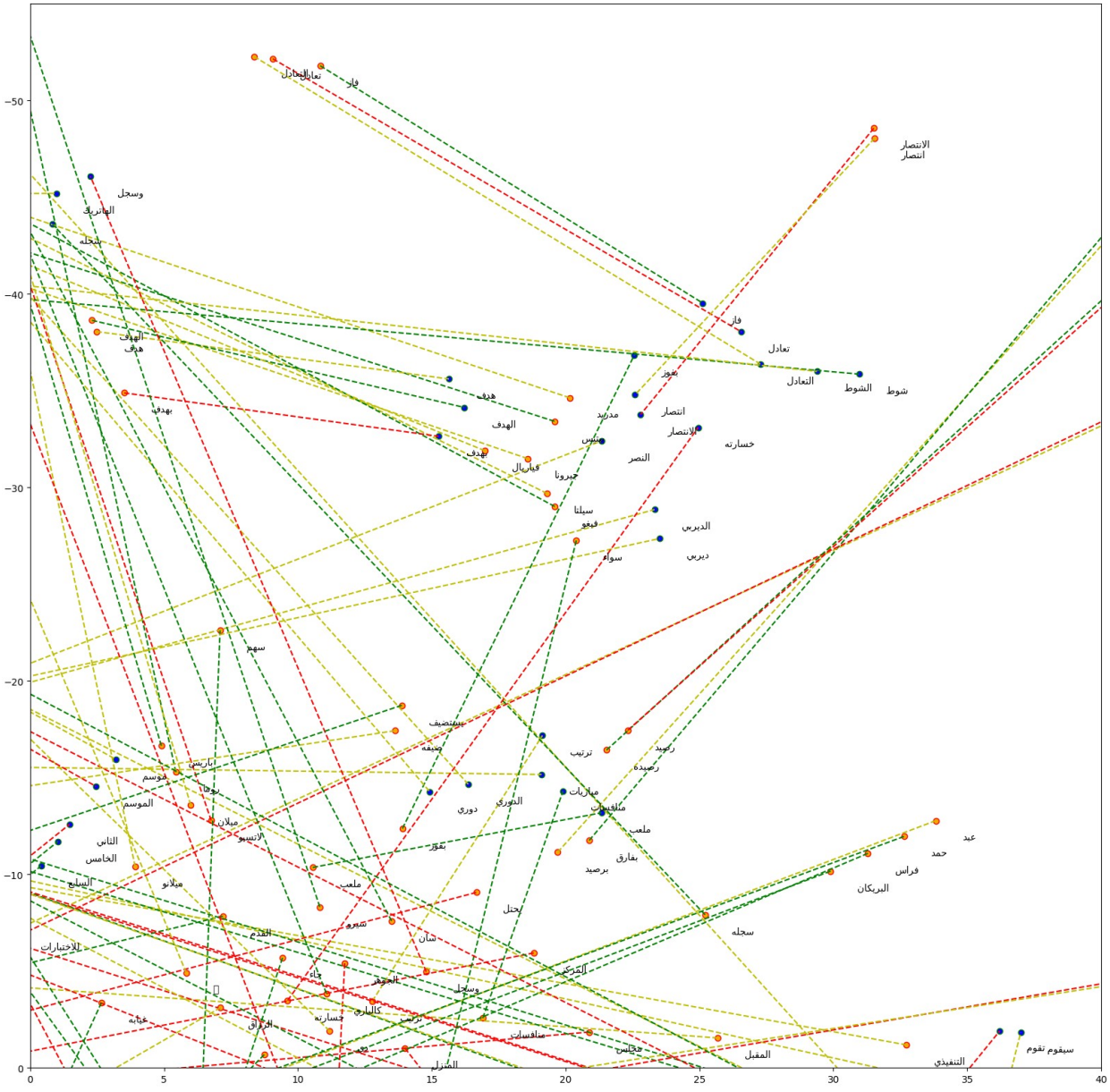
```
from gensim.models import Word2Vec
model_cbow = Word2Vec.load(r'C:\Users\ACER\OneDrive\Documents\MST\S2\
NLP\w2v_CBOW_500_5_400_10.model')

import fasttext
model_fasttext = fasttext.load_model(r'C:\Users\ACER\OneDrive\
Documents\MST\S2\NLP\fasttext arabic\cc.ar.300.bin')

def get_max_value_pair(pair_list):

  if not pair_list:
    return None

  max_pair = pair_list[0]
  for pair in pair_list:
    if pair[1] > max_pair[1]:
      max_pair = pair

  return max_pair

most_sim = {}
for word in filtred_content2:
    if word in model_cbow.wv:
        most_sim[word] =
get_max_value_pair(model_cbow.wv.most_similar(word))
    else:
        most_sim[word] = None
{k:most_sim[k] for k in list(most_sim)[:10]}
```

```
{'أعلن': None,
 'آياكس': None,
 'أمستردام': None,
 '(0.5640104413032532 ,'البرتغالي') :'الهولندي',
 '(0.5962443947792053 ,'لليوم') :'اليوم',
 'الثلاثاء،': None,
 'إيقاف': None,
 'أليكس': None,
 '(0.597771167755127 ,'مودريتش') :'كروس',
 'الرئيس': None}
```

```
vectorized_words_cbow={}
vectorized_words_ft={}
for word in filtred_content2:
    try:
        vectorized_words_cbow[word]=model.wv[word]
        vectorized_words_ft[word] =
```

```
model_fasttext.get_word_vector(word)
    except:
        continue

from sklearn.manifold import TSNE
import  numpy as np
tsne = TSNE(n_components=2,perplexity=2,random_state=0)

reduced_cbow = tsne.fit_transform(np.array([vec.tolist() for vec in
vectorized_words_cbow.values()]))
reduced_ft = tsne.fit_transform(np.array([vec.tolist() for vec in
vectorized_words_ft.values()]))
words = [word for word in vectorized_words_cbow.keys()]

import matplotlib.pyplot as plt
import random
import arabic_reshaper
from bidi.algorithm import get_display

plt.figure(figsize=(20, 20))
plt.scatter(reduced_cbow[:, 0], reduced_cbow[:, 1], c='orange',
edgecolors='r')
plt.scatter(reduced_ft[:, 0], reduced_ft[:, 1], c='blue',
edgecolors='g')
for label, x_cbow, y_cbow, x_ft, y_ft in zip(words, reduced_cbow[:,
0], reduced_cbow[:, 1], reduced_ft[:, 0], reduced_ft[:, 1]):
    label_ = get_display( arabic_reshaper.reshape(label))
    plt.annotate(label_, xy=(x_cbow+1, y_cbow+1), xytext=(0, 0),
textcoords='offset points')
    plt.annotate(label_, xy=(x_ft+1, y_ft+1), xytext=(0, 0),
textcoords='offset points')
    plt.plot([x_cbow, x_ft], [y_cbow, y_ft],
random.choice(['y--','r--','g--']))

xmin, xmax = 0,40
ymin, ymax = 0, -55
plt.xlim(xmin, xmax)
plt.ylim(ymin, ymax)

plt.show()
```

```
C:\Users\ACER\PycharmProjects\LogisticRegression\env\python311\Lib\
site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 65010
(\N{ARABIC LIGATURE ALLAH ISOLATED FORM}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
```



NB: you can change the xmin, xmax, ymin and ymax for zooming on a specific area, remove if no zoom is required

Conclusion:

Based on the plot, it appears that the word embeddings from the CBOW and FastText models for the same words are distinctly different. The lines connecting the same word in both models indicate a significant shift in the vector space. This suggests that each model captures different semantic and syntactic aspects of the words.