

## 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. وأكد آياكس في بيان أن 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 أن "x" في 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 مؤقت. يُذكر أن آياكس يحتل المركز الخامس في ترتيب الدوري الهولندي لكرة القدم برصيد 44 نقطة، بفارق 28 نقطة عن أيندهوفن، المُتصدر.

قدّم روما اليوم الثلاثاء القميص الذي سيرتديه فريق كرة القدم السبت المقبل في مواجهة "ديربي العاصمة" أمام لاتسيو. واستوحى 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 الـ 31 من الدوري الإيطالي. 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 و 75 يورو للأطفال.

أعلن نادي ريال بيتيس اليوم الثلاثاء، إصابة لاعب وسط الفريق، الأرجنتيني إزيكيل أفيلا في أن الفحوصات الطبية التي خضع لها "x" ساقه اليسرى. وأكد ريال بيتيس في تغريدة في منصة 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 المقبل، سيلتا فيغو.

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 والخصوصية والحفاظ على سرية المعلومات". وتابع "من حق المنظمة، وطبقاً للكوند 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 الموضوعّة 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 والخطئة

كشفت نادي الهلال السعودي، أمس الاثنين، طبيعة إصابة مُهاجمه المصري ألكسندر ميتروفيتش. إنَّ الأشعة التي أجراها "x" ومدة 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 أنَّ الهلال يتصدّر ترتيب '، '، \n\n الدوري السعودي برصيد 71 نقطة، بفارق 12 نقطة عن أقرب ملاحقيه، النصر

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 '، '، \n\n صاحب المركز الـ4 الاتحاد

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 لفاريال في الدقيقة الـ51، وسجل ساول نيجوير الهدف الثاني "لاروخي بلانكوس" في الدقيقة الـ88. ورفع أتلتيكو مدريد رصيده بهذا الانتصار إلى '، '، \n\n 58 نقطة في المركز الـ4، وتجمد رصيد فياريال عند النقطة الـ38 نقطة في المركز الـ10 '، '، \n\n

فاز إتنر ميلانو على ضيفه إمبولي، بنتيجة 2-0، اليوم الاثنين، ضمن منافسات الجولة الـ30 من دوري الدرجة الأولى الإيطالي "السيرى آ". أنهى إتنر ميلانو الشوط الأول متقدماً على إمبولي، بهدف سجله الإيطالي فيديريكو ديماركو في الدقيقة الـ6. وفي الشوط الثاني، أضاف أليكسيس سانشير الهدف الثاني "لنيراتزوري" في الدقيقة الـ81، لتبقى النتيجة على حالها حتى صفارة النهاية. ورفع إتنر ميلانو رصيده بهذا الانتصار إلى 79 نقطة في صدارة الترتيب، وتجمد رصيد '، '، \n\n إمبولي عند النقطة الـ25 في المركز الـ18 '، '، \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 المضييفة 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 '، '، \n\n تندرج ضمن شعار غير مسؤول مناهض لروسيا، أو بيان رسمي، مشيراً إلى أن الجوهر الأساسي، وهو "أنهم حقاً لا يريدون رؤية حتى رياضيين حياديين من روسيا في الألعاب الأولمبية '، '، \n\n

تعادل روما مع مضيفه ليتشي، من دون أهداف (0-0)، اليوم الاثنين، ضمن منافسات الجولة الـ30 من دوري الدرجة الأولى الإيطالي، "السيرى 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 '، '، \n\n سيزور نادي ليتشي ملعب "سان سيرو" ليواجه ميلان

فرض التعادل الإيجابي 1-1 نفسه على مباراة أودينيزي مع مُضيفه ساسولو، اليوم الاثنين، ضمن منافسات الجولة الـ30 من الدوري الإيطالي لكرة القدم، "السيرى 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 تلقى 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، والكامبروني 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 '\n\n'."الهاتريك" في انتصار فريقه على الطائي بنتيجة 5-2 في الجولة الماضية

tokenizing the data

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

for article in articles:
    tokenized_articles.append(word_tokenize(article))
tokenized_articles[:10]
```

```
[['أعلن',
  'أياكس',
  'أمستردام',
  'الهولندي',
  'اليوم',
  'الثلاثاء',
  'إيقاف',
  'أليكس',
  'كروس',
  'الرئيس',
  'التنفيذي',
  'الجديد',
  'بشكل',
  'فوري',
  'للاشتباه',
  'في',
  'تورطه',
  'بتداولات',
  'داخلية',
  'الأسهم',
  'النادي.وأكد',
  'أياكس',
  'في',
  'بيان',
```

, 'أَنَّ  
, 'قرار  
, 'إيقاف  
, 'الرئيس  
, 'التنفيذي  
, 'جاء  
, 'بعد  
, 'أَنَّ  
, 'علم  
, 'مجلس  
, 'الإدارة  
, 'أَنَّ  
, 'كروس  
, 'اشترى  
, 'أكثر  
, 'من  
, '17  
, 'ألف  
, 'سهم  
, 'في  
, 'آياكس  
, 'قبل  
, 'أسبوع  
, 'من  
, 'الإعلان  
, 'عن  
, 'تعيينه  
, 'في  
, '2  
, 'آب/أغسطس  
, 'وكشف. 2023  
, 'النادي  
, 'أَنَّ  
, 'الاستشارة  
, 'القانونية  
, 'التي  
, 'أجراها  
, 'أشارت  
, 'إلى  
, 'أنه  
, 'من  
, 'المحتمل  
, 'أَنَّ  
, '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',  
'الكرة',  
'القدم',  
'برصيد',  
'44',  
'،نقطة',  
'بفارق',  
'28',

, 'نقطة'  
, 'عن'  
, 'آيندهوفن'  
, 'المُتصدر'  
, ']'  
[ 'قَدِّم'  
, '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'  
'ال31'  
'من'  
'الدوري'  
'الإيطالي'  
'وتهدف'  
'مبادرة'  
'روما'  
'إلى'  
'إحياء'  
'ذكرى'  
'أحد'  
'أفضل'  
'عصور'  
'النادي.وقال'  
'النادي'  
'في'  
'البيان'  
'إن'  
'القميص'  
'القديم'  
'أعيد'  
'تصميمه'  
'لتكيفه'  
'على'  
'الأذواق'  
'المعاصرة'  
'للجيل'  
'الجديد'  
'من'  
'ال جماهير'  
'وطرح.'



, 'القميص'  
, 'اليوم'  
, 'الثلاثاء'  
, 'للبيع'  
, 'بسعر'  
, '100'  
, 'يورو'  
, 'و75'  
, 'يورو'  
, 'للأطفال'  
, ']'  
, 'أعلن'  
, 'نادي'  
, 'ريال'  
, 'بيتيس'  
, 'اليوم'  
, 'الثلاثاء'  
, 'إصابة'  
, 'لاعب'  
, 'وسط'  
, 'الفريق'  
, 'الأرجنتيني'  
, 'إزكييل'  
, 'أفيلا'  
, 'في'  
, 'ساقه'  
, 'اليسرى.وأكد'  
, 'ريال'  
, 'بيتيس'  
, 'في'  
, 'تغريدة'  
, 'في'  
, 'منصة'  
, 'x'  
, 'أن'  
, 'الفحوصات'  
, 'الطبية'  
, 'التي'  
, 'خضع'  
, 'لها'  
, 'أفيلا'  
, 'أظهرت'  
, 'معاناته'

'من',  
'إصابة',  
'في',  
'العضلات',  
'المأبضية',  
'يساقه',  
'اليسرى.وأصيب',  
'أفيلا',  
'خلال',  
'مباراة',  
'فريقه',  
'مع',  
'جبرونا',  
'في',  
'الجولة',  
'الأخيرة',  
'من',  
'الدوري',  
'الإسباني',  
'والتي',  
'انتهت',  
'بفوز',  
'الفريق',  
'',  
'الكतालوني',  
'',  
'بنتيجة',  
'ويحتل.2-3',  
'ريال',  
'بيتيس',  
'المركز',  
'السابع',  
'في',  
'ترتيب',  
'',  
'الليغا',  
'',  
'برصيد',  
'42',  
'نقطة',  
'ويستضيف',  
'يوم',  
'الجمعة',  
'،المُقبل',  
'سيلتا'

' فيغو ,  
' ] ,  
[ ' 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',  
, 'وتضرر',  
, ' بالرياضيين',  
, ' والرياضة',  
, 'المصرية',  
, 'ومن',  
, 'يخالف',  
, 'ذلك',  
, 'يعرض',  
, 'نفسه',  
, 'للمساءلة',  
, '""',  
, 'وأضاف.',  
, ':",  
, 'لا',  
, 'تقوم',  
, 'المنظمة',  
, 'المصرية',  
, 'لمكافحة',  
, 'المنشطات',  
, 'ينشر',  
, '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 ,  
' , السحب ,  
' , 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  
, 'صباحي.وفي  
, 'وقت  
, 'لاحق  
, 'أعلن  
, 'بيراميدز  
, 'رغبة  
, 'رمضان  
, 'صباحي  
, 'في  
, 'مقابلة  
, 'جمال  
, 'علام  
, ' ,  
, 'بسبب  
, 'التصريحات  
, 'غير  
, 'المسؤولة  
, 'والخاطئة  
, ' ,  
, ' ] .

[ 'كشف ' ,  
'نادي ' ,  
'الهلال ' ,  
'السعودي ' ,  
'أمس ' ,  
'الاثنين ' ,  
'طبيعة ' ,  
'إصابة ' ,  
'مهاجمه ' ,  
'الصربي ' ,  
'ألكسندر ' ,  
'ميتروفيتش ' ,  
'ومدة ' ,  
'غيابه ' ,  
'عن ' ,  
'الملاعب.وقال ' ,  
'الهلال ' ,  
'في ' ,  
'تغريدة ' ,  
'في ' ,  
'منصة ' ,  
'\ ' ,  
'x ' ,  
'\ ' ,  
'إن ' ,  
'الأشعة ' ,  
'التي ' ,  
'أجراها ' ,  
'ميتروفيتش ' ,  
'\ ' ,  
'أكدت ' ,  
'تعرّضه ' ,  
'للإصابة ' ,  
'في ' ,  
'العضلة ' ,  
'الخلفية ' ,  
'وسخّض ' ,  
'على ' ,  
'إثرها ' ,  
'البرنامج ' ,  
'علاجي ' ,  
'وتأهيلي ' ,  
'لمدة ' ,  
'6 ' ,  
'أسابيع ' ,  
'\ ' ,



, 'وأصيب .  
, 'ميتروفيتش  
, 'خلال  
, 'مباراة  
, 'فريقه  
, 'أمام  
, 'الشباب  
, 'في  
, 'الجولة  
, 'الماضية  
, 'من  
, 'دوري  
, 'روشن  
, 'السعودي  
, 'والتي  
, 'انتهت  
, 'بفوز  
,  
, 'الزعيم  
,  
, 'بنتيجة  
, 'يُذكر. 3-4  
, 'أنَّ  
, 'الهلال  
, 'يتصدّر  
, 'ترتيب  
, 'الدوري  
, 'السعودي  
, 'برصيد  
, '71  
, 'نقطة  
, 'بفارق  
, '12  
, 'نقطة  
, 'عن  
, 'أقرب  
, 'ملاحقيه  
, 'النصر  
, '],  
[ 'فاز'  
, 'الأهلي'  
, 'على'  
, 'ضيفه'  
, 'الاتحاد'  
, 'بنتيجة'

1-0 ،  
'الاثني ،  
'ضمن ،  
'منافسات ،  
'الجولة ،  
'ال26 ،  
'من ،  
'دوري ،  
'روشن ،  
'السعودي ،  
'لكرة ،  
'القدم. سجل ،  
'هدف ،  
'المباراة ،  
'الوحيد ،  
'لنادي ،  
'الأهلي ،  
'فراس ،  
'البريكان ،  
'في ،  
'الدقيقة ،  
'ال34 ،  
'لتبقى ،  
'النتيجة ،  
'على ،  
'حالتها ،  
'حتى ،  
'صفارة ،  
'النهاية ،  
'ويحقق ،  
'الملكلي ،  
'" " ،  
'العلامة ،  
'الكاملة. ورفع ،  
'الأهلي ،  
'رصيده ،  
'بهذا ،  
'الانتصار ،  
'إلى ،  
'51 ،  
'نقطة ،  
'في ،  
'المركز ،  
'ال3

'وبفارق',  
'5',  
'نقاط',  
'عن',  
'صاحب',  
'المركز',  
'ال4',  
'الاتحاد',  
'[',  
'فاز',  
'أتلتيكو',  
'مدريد',  
'على',  
'مضيفه',  
'، فياريال',  
'بنتيجة',  
'2-1',  
'اليوم',  
'، الاثنين',  
'ضمن',  
'منافسات',  
'الجولة',  
'ال30',  
'من',  
'دوري',  
'الدرجة',  
'الأولى',  
'الإسباني',  
'',  
'الليغا',  
'"',  
'أنهى',  
'أتلتيكو',  
'مدريد',  
'شوط',  
'المباراة',  
'الأول',  
'متقدماً',  
'على',  
'، فياريال',  
'بهدف',  
'سجله',  
'اللاعب',  
'البلجيكي',  
'أكسيل'

'فيتسل',  
'في',  
'الدقيقة',  
'ال9.وفي',  
'الشوط',  
'الثاني',  
'أدرك',  
'المهاجم',  
'النرويجي',  
'ألكسندر',  
'سورلوث',  
'التعادل',  
'لفياربال',  
'في',  
'الدقيقة',  
'ال51',  
'وسجل',  
'ساول',  
'نيغوينز',  
'الهدف',  
'الثاني',  
'',  
'للروخي',  
'بلانكوس',  
'',  
'في',  
'الدقيقة',  
'ال88.ورفع',  
'ألتتيكو',  
'مدريد',  
'رصيده',  
'بهذا',  
'الانتصار',  
'إلى',  
'58',  
'نقطة',  
'في',  
'المركز',  
'ال4',  
'وتجمد',  
'رصيد',  
'فياربال',  
'عند',  
'النقطة',  
'ال38',

, 'نقطة'  
, 'في'  
, 'المركز'  
, 'ال10'  
, ']' ,  
, ' فاز '[  
, 'إنتر'  
, 'ميلانو'  
, 'على'  
, 'ضيفه'  
, 'إمبولي'  
, 'بنتيجة'  
, '2-0'  
, 'اليوم'  
, 'الاثنين'  
, 'ضمن'  
, 'منافسات'  
, 'الجولة'  
, 'ال30'  
, 'من'  
, 'دوري'  
, 'الدرجة'  
, 'الأولى'  
, 'الإيطالي'  
,  
, 'السير'  
, 'أ'  
, ' ' ' ' ,  
, 'أنهى.'  
, 'إنتر'  
, 'ميلانو'  
, 'الشوط'  
, 'الأول'  
, 'متقدماً'  
, 'على'  
, 'إمبولي'  
, 'بهدف'  
, 'سجله'  
, 'الإيطالي'  
, 'فيدريكو'  
, 'ديماركو'  
, 'في'  
, 'الدقيقة'  
, 'ال6. وفي'  
, 'الشوط'

' , الثاني ,  
' , أضاف ,  
' , أليكسيس ,  
' , سانشيز ,  
' , الهدف ,  
' , الثاني ,  
' ,  
' , للنيراتزوري ,  
' ,  
' , في ,  
' , الدقيقة ,  
' , الـ 81 ,  
' , ليتقى ,  
' , النتيجة ,  
' , على ,  
' , حالها ,  
' , حتى ,  
' , صفارة ,  
' , النهاية. ورفع ,  
' , إنتر ,  
' , ميلانو ,  
' , رصيده ,  
' , بهذا ,  
' , الانتصار ,  
' , إلى ,  
' , 79 ,  
' , نقطة ,  
' , في ,  
' , صدارة ,  
' , الترتيب ,  
' , وتجمد ,  
' , رصيد ,  
' , إمبولي ,  
' , عند ,  
' , النقطة ,  
' , الـ 25 ,  
' , في ,  
' , المركز ,  
' , الـ 18 ] ,  
' , انتقد [ ,  
' , رئيس ,  
' , اللجنة ,  
' , الأولمبية ,  
' , الروسية ,  
' , ستانيسلاف ,

' , 'بوزدنياكوف ,  
' , 'اليوم ,  
' , 'الاثنين ,  
' , 'تصريحات ,  
' , 'عمدة ,  
' , 'باريس ,  
' , 'آن ,  
' , 'هيدالغو ,  
' , 'والمعلقة ,  
' , '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',  
'المضيقة',  
'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',  
'تندرج',  
'ضمن',  
'شعار',  
'غير',  
'مسؤول',



'مناهض'  
'لروسيا'  
'أو'  
'بيان'  
'رسمي'  
'مشيراً'  
'إلى'  
'أن'  
'هذا'  
'لا'  
'يغير'  
'الجوهر'  
'الأساسي'  
'وهو'  
'أنهم'  
'حقاً'  
'لا'  
'يريدون'  
'رؤية'  
'حتى'  
'رياضيين'  
'حياديين'  
'من'  
'روسيا'  
'في'  
'الألعاب'  
'الأولمبية'  
'[  
'تعاذل'  
'روما'  
'مع'  
'مضيفه'  
'ليتشي'  
'من'  
'دون'  
'أهداف'  
'( '  
'0-0'  
' )'  
'،'  
'اليوم'  
'، الاثنين'  
'ضمن'  
'منافسات'

'الـجولة',  
'الـ30',  
'من',  
'دوري',  
'الدرجة',  
'الأولى',  
'الإيطالي',  
'',  
'السيرى',  
'أ',  
'''',  
'ورفع.',  
'روما',  
'رصيده',  
'بهذا',  
'التعادل',  
'إلى',  
'52',  
'نقطة',  
'في',  
'المركز',  
'الـ5',  
'وبات',  
'رصيد',  
'ليتشي',  
'29',  
'نقطة',  
'في',  
'المركز',  
'الـ13.وسيوخوز',  
'فريق',  
'المدرّب',  
'دانييل',  
'دي',  
'روسي',  
'',  
'ديريبي',  
'''',  
'العاصمة',  
'الإيطالية',  
'في',  
'الـجولة',  
'المقبلة',  
'عندما',  
'يستضيف',  
'لاتسيو',

```
'في',  
'حين',  
'سيزور',  
'نادي',  
'ليتشي',  
'ملعب',  
,,  
'سان',  
'سيرو',  
"',  
'ليواجه',  
'ميلان',  
['.']]
```

\_\_\_\_\_

```
filtered_content = [word for article in tokenized_articles for word in
```

```
print('before filtration: ', len([word for article in
```

```
print('after filtration: ', len(filtered_content))
```

\_\_\_\_\_

before filtration: 1215  
 filter: 888

91101 - 11/11/2017 - 500

```
nettover = re.compile(r'\d\d')
```

```
digit_words = [word_string for w, i in filtered_content for word in
```

```
def _finditer(pattern, string, flags):
    """Return an iterator over all non-overlapping matches
    of the pattern in the string. The matches are returned as
    Match objects.

    Optional flags can be passed to the function, for example
    re.IGNORECASE or re.DOTALL.

    Stop iteration when the pattern is no longer found.
    """
    return _iter_re_finditer(pattern, string, flags)
```

[illegible]

```
filtered_content2 = [word for word in filtered_content if not word in
```

```
print('before filtration: ', len(filtered_content))
```

```
print('before filtration: ', len(filtered_content))
print('after filtration: ', len(filtered_content))
```

```
print('direct: ', direct, '\n', sep='')
print('indirect: ', indirect, '\n', sep='')
```

before filtration	550
after filtration	667

after: 11/11/2011 00:

91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000

```
one_hot_encoded_data = pd.get_dummies(filtered_content)
```

```
one_not_encoded_data = part.get_admin_id(1)
one_not_encoded_data
```

— — — — —

	-							\ أنهى .يادر .وأشارت .وأصيب .وأضاف .وتابع .ورفع .وطرح .
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

\ وکان ... يحتل يحق يخالف يريدون يستضيف يعرض يغير يفيد.									
0	False	...	False	False	False	False	False	False	False
1	False	...	False	False	False	False	False	False	False
2	False	...	False	False	False	False	False	False	False
3	False	...	False	False	False	False	False	False	False
4	False	...	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
862	False	...	False	False	False	False	False	False	False
863	False	...	False	False	False	False	False	False	False
864	False	...	False	False	False	False	False	False	False
865	False	...	False	False	False	False	False	False	False
866	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 filtered_content2:
    if word not in word2count.keys():
        word2count[word] = 1
    else:
        word2count[word] += 1
{k:word2count[k] for k in list(word2count)[:10]}

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

```

TF-IDF

```

from sklearn.feature_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer()
result = tfidf.fit_transform(filtered_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())

```

## Applying 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 filtered_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,
 'الهولندي': ('البرتغالي', 0.5640104413032532),
 'اليوم': ('اليوم', 0.5962443947792053),
 'الثلاثاء': None,
 'إيقاف': None,
```

```

'أليكس': None,
'كروس': ('مودريتش', 0.597771167755127),
'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_resaper
from bidi.algorithm import get_display

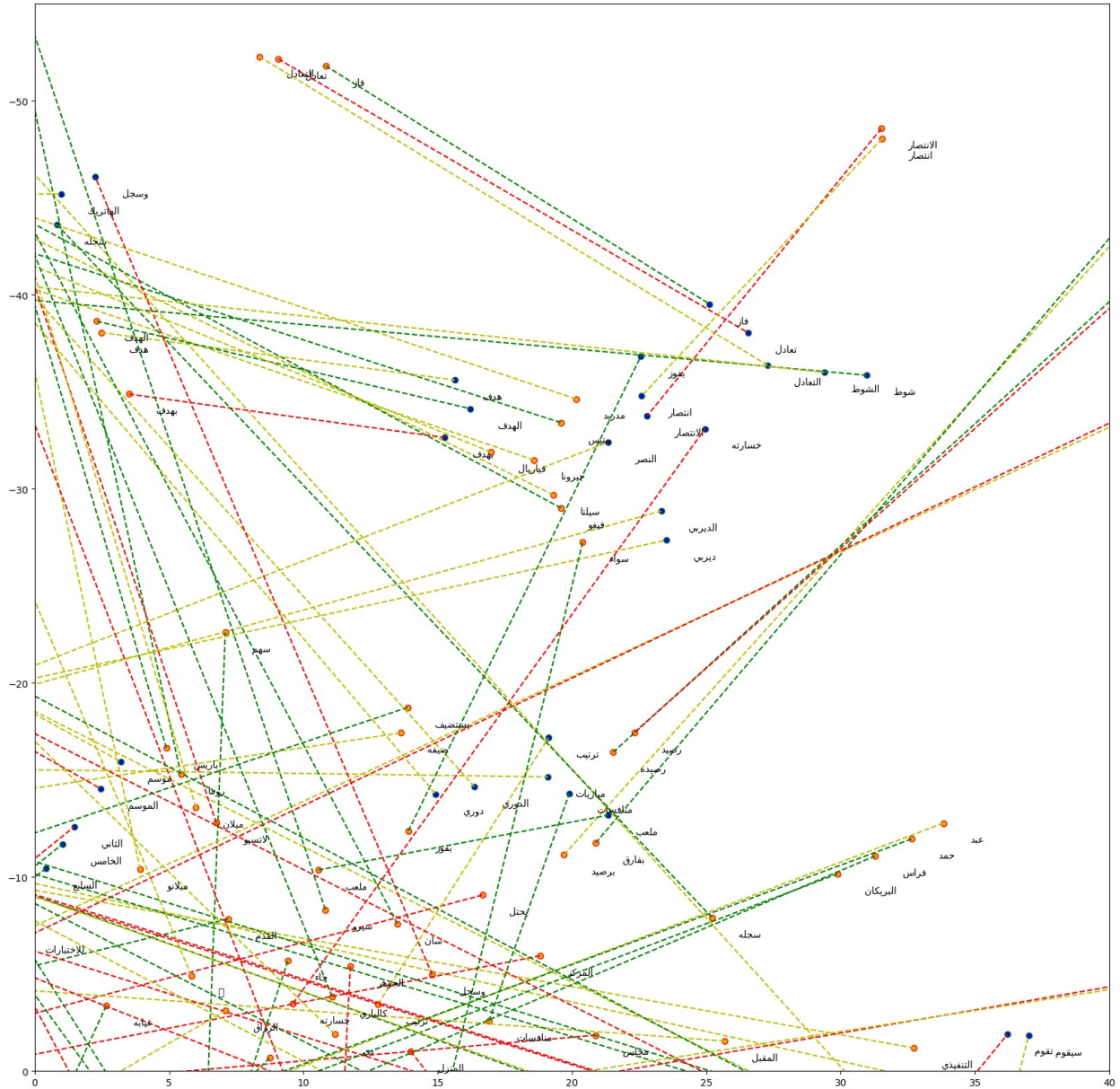
plt.figure(figsize=(20, 20))
plt.scatter(reduced_cbow[:, 0], reduced_cbow[:, 1], c='orange',
edgecolors='r')
plt.scatter(reduced_ft[:, 0], reduced_ft[:, 1], c='blue',
edgecolors='g')
for label, x_cbow, y_cbow, x_ft, y_ft in zip(words, reduced_cbow[:,
0], reduced_cbow[:, 1], reduced_ft[:, 0], reduced_ft[:, 1]):
    label_ = get_display( arabic_resaper.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.

