Post 1

الموديل فى الاخر بيحتاج منى انى ادخلة أرقام ، الأرقام ديه هو من خلالها يقدر يشوف ال pattern من خلالها بطريقة ما ، فلما نيجى ننقل لل NLP هلاقى الارقام ديه او ال Features مبقتش متوفرة علطول عشان اديها لل model فهو يقدر يربط الدنيا ببعض ويفهم ، بقيت أنا محتاج انى دلوقت اعمل ال Features وديه ومش لاى ارقام لا لارقام هو فعلا يقدر يلاقى خلالها pattern فيفهم انا محتاج ققولة ايه من خلال الارقام ديه فبقى عندى مرحلتين مهمين بالنسبة ليا قبل ما ال model بتاعى يشتغل ، المرحلة الأولى هى Preprocessing لل Text بتاعى ، سواء انى اعمله cleaning او اشيل ال Stop words او اعمل Lemmmization وغيرة من الخطوات الى بظبط بيها ال Text بتاعى فى Form معينة تتناسب مع ال Application بتاعى لان بعض الخطوات ممكن تكون متضمنة فى Application ما واخر لا ، وكمان لأن بعض الخطوات ديه او معظمها بيكون ليه بعض الإيجابيات وبعض السلبيات ، بعد كده بتيجى المرحلة الى فيها تحول ال Text ده بقا لأرقام وهى ال Preparation الى بقدر من خلالها ادى ال Data بتاعتى الى كانت مجرد Text من لغة ما لأرقام الموديل يقدر يشتغل معاها ويطلع نتائج ، المرحلة الأولى ديه فيها بعض الخطوات المهمة جدا والى من خلالها أصلا هقدر انى اتعامل مع معظم الخطوات الى بعدها وهى ال Tokenization وده نتكلم عنه بإذن الله البوست الجاى

Post 2

ال Tokenization هو إنى اعمل Segmentation للكلام الى عندى يعنى لو عندى paragraph احولى لمجموعة من الجمل ، والجمل ديه احولها لكلمات ، فى الاخر انا هحتاج انى احول ل Basic level الى بيكون الجملة وهو الكلمة ، بعد كدة ببتدى أشوف الكلمات ال Unique فى الداتا الى عندى واستخدمها فى انى احولها لأرقام بناء على بعض الطرق هنجلها بعدين والارقام ديه الى هتاخدها الكلمات هبتدى اعوض بيها مكان الكلمة فى كل جملة وبرضه ده هنعرفه قدام بإذن الله ، لكن نرجع تانى لعملية ال Tokenization لانها بتختلف بإختلاف اللغة وال Application الى بتعامل معاه ، مثلا هل كلمة زى dont هعتبر كلمة واحد ولا هحتاج انى احولها ل do and not طيب كلمة زى Ice cream هل هعتبرها ك Token واحد ولا اتنين ؟ طيب هل الكلمة ديه ليها Predefined rule فى اللغة او معنى معين زى مثلا dont ديه بتعبر عندى عن حالة نفى ، كمان الكلمات نفسها و أشكالها المختلفة زى كمة end and ending هل كل كلمة من دول هعتبرها كلمة لوحدها ولا محتاج ارجع الكلمة لأصلها زى ما هنشوف فى ال Steaming and lemmization ولكن نلاقى مثلا كلمة زى run and running لو هرجعها ل run ك أصل الكلمة محتاج اشيل ning بينما فى ending محتاج اشيل ال ing فقط ولو كلمة زى bus هل ال s فى الاخر هى suffix زى ما فى كلمة words كل ديه cases وغيرها محتاج اعرفها وانا شغال وهتعامل معاها ازاى لانها هتتسبب فى تكبير الكلمات بتاعتى او ما يسمى Bag-of-words ، فى الاخر ال Tokenization هو تقسيم الجملة او الكلام لمجموعة من الكلمات ولكن محتاج اعرف التقسيم ده هيكون بناء على ايه ، وهنا ال Libraires المختلفة منها الى بيستخدم rules من اللغة ومنها الى شغال عن طريق انه يعمل split عن طريق ال space وممكن انتا تعمل ال Tokenization الخاص بيك بناء على Regular expression معين او rules معينة بإستخدام Regex or re libraires .

```
In [2]: sentence = """Thomas Jeferson began building Monticello at the age at 26."""

        sentence.split()

Out[2]: ['Thomas',
         'Jeferson',
         'began',
         'building',
         'Monticello',
         'at',
         'the',
         'age',
         'at',
         '26.']
```

Post 3

من اول stage ال nlp هلاقى فيه عندى challenges كتير لأن كل حاجه ممكن يكون ليها منفعة ما ممكن تضرنى بشكل اخر ، وهنا بحتاج اشوف كل خطوة وارتباطها بالابليكشن الى انا شغال فيه وازاى هتاثر فيها بشكل او باخر فمحتاج اشوف مثلا الفرق بين انى اعمل Steaming and lemmitazation واخد فى اعتبارى انى فيه بعض الكلمات بسببهم ممكن ميكونش ليها اى

معنى واصبحت اصلا ملهاش علاقة باللغة وبذات فى حالة ال steaming لانه بيشيل ال suffix and prefix فكلمة مثلا زى ending هيشيل منها ing طب بالنسبة لكلمة زى sign لو شيلنا ال ing بقت مجرد حرف s وكلمة زى bus and words لما اشيل ال s اصبحت ليها معنى مع words لكن بالنسبة ل bus بقت bu وكلمة زى believe لما اعمل steaming هتكون believ برضه ملهاش اى معنى فى اللغة ، بينما ال Lemmization بيكون ليها علاقة بال Grammers بتاع اللغة وبياخد في اعتبارة الأشكال المختلفة للكلمة بناء على السياق الى جت فيه ، وبعض المكتبات بتحدد فيها الأولوية لل verb or noun لما يكون عنده اكتر من اختيار للكلمة ، والمهم فى الخطوتين دول انهم بيساعدوا ال model بعد كده انه ميحصلش فيه over fitting لانه بدل ما كان بيدى attention لكل أشكال الكلمة لا دلوقت بقى عندة شكل او اتنين منها فى ال vocab الى انتا بتبنيها وده بيقلل ال over fitting لانه لو الكلمة كانت جت فى شكل معين مرة واحدة فى كل ال data هيديها اهمية لانك هتعتبرها من ضمن ال vocab فمش هيقدر يعمل generalization وبجانب ده كمان انتا بتقلل ال Vocab بتاعتك الى بدوره بيقلل معاك ال computation الى بتحتاجها ، طب ايه هى ال Vocab ديه وازاى ببنيها هنشوفها فى البوست الجاى بإذن الله

Post 4

بعد ما عملنا Tokenization لل Text وحددنا لو هنعمل Lemmtization or stemming بنبتدى ننقل من مرحلة ال Pre-processing text لمرحلة انى اعمل Pre-paration لل text وده وهو انى احول الكلام ده لارقام الموديل يقدر يفهمها ، واحدة من الطرق ديه هى ال One-hot encoder وهو انى مجرد ما خلاص حددت ال (Vocabulary (nuique words الى هستخدمها هبتدى بقا انى اعمل matrix n*V ، لكل document عدد ال Token الى فيه فى عدد Vocabulary يعنى لو كان document من الداتا مثلا 100 كلمة وعدد الكلمات الى فى الداتا كلها 20000 فهنا هحتاج matrix 100*20000 فقط ل document واحد، والحقيقة انى ده رغم انه مكلف جدا جدا وفى الحياه العملية مينفعش يشتغل لان اللغة نفسها natural ولو اعتبرنا كل كلمات اللغة وال symbols وغيره من الحاجات المتضمنه من اللغة هيكون مستحيل اعمل process وهتكون عملية مكلفة جدا وفى الاخر بيكون عندى sparse matrix كلها اصفار وغير ، الا ان ال one-hot من خلالة بقدر احافظ على ال context and grammar that words comes in فالموديل هيقدر ياخد بالة من ال sequence بتاع الكلمات لكنه طبعا لو قلنا عندنا 3000 كتاب وكل كتاب فية حوالى 3500 جملة وكل جملة 15 كلمة ف احنا محتاجين شوف محتاجين بقا مساحة قد ايه غير انى ده كلة هيكون مضروب فى عدد ال Vocabulary بتاعتك ، بعد كده ابتدينا نتجاهل ال order عشان ننتقل لفكرة ال Bag of words وده فى البوست الجاى.

```
Thomas Jeferson began building Monticello at the age at 26.
==================================================

Out[4]: array([[0, 0, 0, 1, 0, 0, 0, 0, 0],
               [0, 1, 0, 0, 0, 0, 0, 0, 0],
               [0, 0, 0, 0, 0, 0, 1, 0, 0],
               [0, 0, 0, 0, 0, 0, 0, 1, 0],
               [0, 0, 1, 0, 0, 0, 0, 0, 0],
               [0, 0, 0, 0, 0, 1, 0, 0, 0],
               [0, 0, 0, 0, 0, 0, 0, 0, 1],
               [0, 0, 0, 0, 1, 0, 0, 0, 0],
               [0, 0, 0, 0, 0, 1, 0, 0, 0],
               [1, 0, 0, 0, 0, 0, 0, 0, 0]])
```

Post 5

ال Bow of words هو فكرة بتعتمد على ال Vocabulary الى فى الداتا ال unique words وليه طرق مختلفة من binary to counts to frequency and lastly the TF-idf ، بعد فكرة ال one-hot الى كانت بتحافظ على ال sequence بتاعى لكنها كانت استحالة نقدر نتعامل بيها لانها مكلفة جدا جدا ومحتاجة resource كبيرة لداتا ممكن تكون قليلة جدا بالنسبة للداتا الحالية الى بنتعامل معاها ، فقلنا لو تجاهلنا ال order بتاع الكلمات وروحنا عملنا sum لل columns هنلاقى بقا عندنا one vector بعدد ال vocabulary عبارة عن 0 و 1 لو كانت الكلمة ديه جت هنلاقي نتيجة ال column ده 1 لو مكنتش جت فى ال document الى انا فيه هنلاقى 0 ولكن هنا مبقاش فيه order وممكن علاجة حاجة زى كده هى تقسيم ال

document نفسه ل sentence وال ml فى التعلم هيقدر برضه يحصل على معلومات لان الجملة فى الاخر عبارة عن كام كلمة لو حتى ال order بتاعها اتلغبط احنا نفسنا بنقدر نفهم بتتكلم عن ايه ونرتبها كمان بظبط زى ما بتيجى تتعلم لغة جديدة ، بعد كده بدل ال binary representation ده بقينا نعبر بال counts عشان يكون فيه different weights والموديل يقدر ياخد بالة من الكلمات الى عدد تكرارها اكتر لكنه بقا فيه مشكلة وهو انى الكلمات الى استخدامها شائع جدا فى document هتاخد weights اكبر ، ومشكلة اخر فى ال document الكبيرة هنلاقى بعض الكلمات خددت counts كبير جدا والبعض الاخر واخد counts قليل جدا وبقا فيه ranges مختلفة فرجعنا لعملية ال normalization وهو انى اعمل normalize للكلمات بالنسبة لل Vocubalry فبقيت فى range من 0 ل 1 ولكن ما زال ده كل مجرد counts ومش بيعبر عن لا معانى الكلمات او حتى اهميتها وبالنسبة لل rare words الكلمات الاقل شيوعا الى بتكون بالنسبه document ما هى الاثاث هنلاقى ال weights بتاعتها لا تذكر وبالتالى ال model مش هيقدر ياخد بالة منها وده كان ال frequency بتاعت الكلمات، ولكن بعد كده ابتدت ظهور ال TF-idf وده الى هنتكلم عنه المرة الجاية بإذن الله .

Out[10]:

	Thomas	Jeferson	began	building	Monticello	at	the	age	26.	Construction	...	South	Pavilion	in	1770.	Turning	a	neoclassical	masterpiece
sent_1	1	1	1	1	1	1	1	1	1	0	...	0	0	0	0	0	0	0	0
sent_2	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0
sent_3	0	0	0	0	0	0	1	0	0	0	...	1	1	1	1	0	0	0	0
sent_4	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	1	1	1	1

4 rows × 31 columns

Post 6

ال Tokenization بتاعى دوما بيبقا محتاج improvement وبيختلف بإختلاف ال application الى انا شغال فيه ، ساعات بحتاج اعتبر ال punctuations ديه tokens وفى اوقات تانى لا فمثلا لو بنتكلم فى tweets فال symbols معظمها بيكون مهم بالنسبة ليا لانها بتؤدى لمعنى ما زى ال emotions وفى نفس الوقت ممكن ال application نفسه يكون مبنى على tweets لكن مش مهم فيه ال emotions مثلا انى اعتبرها tokens لو انا مثلا شغال Named Entity Recognition ، والتعامل مع ال Tokenization بيختلف بإستخدام ال library الى بتتعامل معاها وال Tokenizer نفسه مثلا NLTK tree bank tokenizer بيشتغل بناء على Grammar rules زى spacy لكن spacy بترجع spacy token ومحتاج اخد بالى من حاجه زى كده واحولها ل strings عشان لما اجى اتعامل معاها زى ال sparse matrix مع tf-idf بحتاج احولة ل array لما اجى اديه للموديل ، وفيه casual tokenizer وده فى حالة انك بتتعامل مع Tweets فبياخد باله من معظم ال symbols ، وفى نفس الوقت انتا ممكن بإستخدام ال Regular expression تعمل implementation لل Tokenizer بناء على regs expression معين وطبعا كل حاجه من ديه بتختلف فى دقتها وسرعتها وده مثلا فرق بين nltk and spacy ل 10 تويتات، كمان ساعات بحتاج يكون ال token نفسة يكون عبارة عن كلمتين او تلاته لانه بيمثل على بعض expression او intuition ما زى ice cream ولكن ده تاثيرة بيكون سيء لو اعتبرت كل ال grams-2 لان معظم الكلمات مش بتيجى في sequence مرتبط ببعضه زى ال ice cream وكل ما ال n-grams بتاعتك بتكبر بيكون ال statics بتاعت حدوث ال sequence ده من الكلمات أقل ، وكمان ده بيساعدنا اننا نحصل على جزء حتى لو بسيط من ال squence ال order الى تجاهلنا من بعد ال one-hot ، المشلكة هنا زى ما قلنا ال statics بتاعت حدوث ال n-grams ده مع بعض بتاثر على الموديل بتاعى وبيحصل overfitting لان الموديل هيبتدى ياخد بالة من sequence هو اصلا بيحدث نادرا فلما يجى فى ال testing هنلاقى نتيجة سيئة ،

```
The number of Tokens are:  188
################################################
The number of Vocabulary are:  133
  vocabulary
0      12398
1         59
2    اتفاقيه
3       احنه
4      استاذه
5       اصابه
6        اصل
7     اكتشاف
8   الاتصالات
9    الاطفال
```

```
The number of Tokens are:  192
################################################
The number of Vocabulary are:  134
  vocabulary
0
1       12398
2          59
3     اتفاقيه
4       احنه
5      استاذه
6       اما به
7        امل
8     اكتشاف
9   الاتصالات
```

Post 7

بعد موضوع ال n-grams هنلاقى حصل زى conflict بين ال n-grams عموما وكل ما تكبر فى ال n ديه بيحصل انى نادر ما ال sequence ده فعلا بيجى مع بعض ، وفى نفس الوقت لو شوفنا ال one-gram هنلاقى انى فيه كلمات rare يعنى ذكرها قليل جدا لكنها مهمه جدا جدا بالنسبة لل document بتاعى ، بينما ال rare sequence فى n-grams هو كده كده مش مهم بالنسبة ليا ، وال n-grams ده كل ما كان قريب من عدد ال docs او اكتر من ال documents of dataset بيحصل overfitting وحل حاجه زى كده بيكون عن طريق انى اظبط مثلا threshold معين لل grams الى هعتبرها زى بظبط موضوع ال stopwords بانى لو كلمة تعدت نسبة معينة هبتدى اتاجهلها وده بيكون بناء على نسبتها بالنسبة لل document الى هى فيه مع ال documents التانية وهو ده الى بيعمله ال Tf-idf انه بيحافظ على rare words وبيديها weights كبير بينما بالنسبة لل stopwords بنلاقيها واخده weights قليلة جدا ، ورغم انى عدد ال stopwords مش بيكون مهم اشيله او اسيبه لانه بالنسبة لل vocabulary size هنلاقى الفرق بسيط وكمان ساعات بيكون ليه اهمية بذات في كلمات زى as or and from والكلمات الى زى كده مع حاجات three-grams زى work at home هنلاقى كلمة at هنا ليها تاثير كبير ، وبرضة من الحاجات المهمة هو ال normalization وهو انى اخلى ال text يكون ليه form واحدة وبينفع مع بعض ال application برضه على عكس application تانية زى موضوع التشكيل بالنسبة للغة زى العربى ساعات كتير بيغير المعنى لكنه بيوفر computation power كبير جدا بسبب انى الكلمة الواحدة ليها تشكيلات مختلفة لكن ممكن فى classifcation problem ده ميكونش مهم او فى حاجة زى Information Retrivers. كل الحاجات ديه وغيرها بتساعد فى تقليل ال vocabulary الى عندى ، وبذات حاجه زى ال stemming and lemmitaization وده الى هنجيلة بإذن الله البوست الجاى .

Post 8

الكلمات فى اللغة بيكون ليها inflections مختلفة أشكال مختلفة سواء فيه prefix or suffix او سواء تشكيل الكلمات زى العربى والتعامل مع ال inflections ديه بيكون عن طريق stemming or lemmitaization ال stemming بيرجع الكلمة لل root بتاعها وبيشيل ال prefix وال suffix للكلمات زى كلمات ending, ends be end لكن لو كلمة زى running هحتاج اشيل ning وكلمة زى sing لو شيلت ing هيتبقى s وغيره لذلك بلاقى كلمات كتيرة مبقاش ليها معنى فى اللغة اصلا ، على عكس ال lemmitaization بيتعامل مع قواعد اللغة والمحتوى الى الكلمة جت فيه + اعتبار الشكل المناسب مع السياق من المعانى المختلفة الى الكلمة ممكن تيجى فيه من فعل او اسم او غيره ، لذلك فيه parameter مهم وهو ال Part of speech موجود لما بنستخدم nltk lemmatizer وهو pos عشان احدد لما يكون عندى اكتر من معنى للكلمة الافضلية تكون فى استخدام الفعل ولا الاسم ولا الصفة وغيرة ، ورغم انى ال stemming بيؤدى ساعات لكلمات مش من اللغة الا انه اسرع من ال lemmitaization .