Activités du Lundi 18 février au Vendredi 22 février :

Fait :

- Vérifier dans quels documents nous pouvons trouver des events avec plusieurs makeinstances
- Récupérer pour chaque event/timex son contexte (4 pos d'avant / 4 pos d'après) (**en cours**)
- Ajouter les 4 mots qui précèdent l'event et 4 mots qui suivent l'event dans le dataframe (**en cours**)
- Prendre en compte le POS tagging (première lettre) pour le lemmatiseur de NLTK (lemmatizer.lemmatize(event, pos='v')
- Nouveau csv pour les signaux

Constats

Events sans makeinstance

En effectuant la tâche de vérification si un event pouvait avoir plusieurs makeinstances, j'ai découvert que dans notre corpus nous n'avons pas ce cas-là. Cependant, je me suis aperçue qu'un event pouvait ne pas avoir de makeinstance.

Les fichiers / events concernés :

AP900816-0139 notified
APW19980213.1310 save
APW19980213.1380 discover
APW19980213.1380 dismembered
APW19980213.1380 kidnapped
ea980120.1830.0071 believe
ea980120.1830.0071 enthusiastic
PRI19980303.2000.2550 has
WSJ900813-0157 trying
WSJ900813-0157 see
WSJ910225-0066 battle
WSJ910225-0066 liberate
WSJ910225-0066 proceeding
WSJ910225-0066 war
WSJ910225-0066 begin
WSJ910225-0066 winding
WSJ910225-0066 offensive
WSJ910225-0066 met
WSJ910225-0066 resistance
WSJ910225-0066 penetrated
WSJ910225-0066 said
WSJ910225-0066 reports
WSJ910225-0066 on
WSJ910225-0066 think
wsj_0124 chairman
wsj_0144 earnings
wsj_0159 offer
wsj_0167 loss
wsj_0176 payments
wsj_0344 sales
wsj_0568 payments
wsj_0568 said
wsj_0583 acquired
wsj_0583 said
wsj_0586 said
wsj_0586 take
wsj_0662 transaction
wsj_0768 redemptions
wsj_0786 offer
wsj_0938 offer

Identifiants artificiels des timex (dataframe_id.csv / dataframe_id_ponctuation.csv)
Ensuite, en voulant assigner un identifiant artificiel aux timex dans le csv « dataframe_id.csv », j'ai procédé à un matching de termes comme pour assigner un identifiant artificiel aux events. Sauf que pour les timex cette méthode ne fonctionne pas :

Ma liste de mots (words) est splité en mots donc : ['bombings', 'in', 'Kenya', 'and', 'Tanzania', 'last', 'week']

Ma liste de timex est splité en timex : ['last week', 'Friday', …]

Lors de la comparaison des termes, je compare un mot (de la colonne word) avec un autre mot (de la liste de timex) : 'last week' ne sera pas détecté, 'Friday' sera détecté. Par conséquent, les colonnes timex et id timex sont peu remplies.

Ce problème devrait être résolu d'ici la semaine prochaine. Anne-Lyse m'a proposé une solution que je vais tenter d'appliquer.

DataFrame de Pandas

Cette semaine j'ai pu découvrir le module DataFrame de la librairie Pandas. J'ai actuellement 3 fichiers csv :

- Dataframe_id (créé avec un DataFrame),
- Csv_features_events,
- Csv_features_timex (Issus du premier script Python avec une structure en dictionnaires de dictionnaires).

Anne-Lyse m'a soumis l'idée de fusionner les lignes de Csv_features_events et Csv_features_timex dans dataframe_id grâce à une clef commune qui permettra de bien aligner le contenu. Cette fusion peut être effectuée avec la fonction 'pd.merge(colonnes)' et le paramètre 'on='clef_commune' (https://www.shanelynn.ie/merge-join-dataframes-python-pandas-index-1/).

Structures des données

En essayant d'adapter ma structure de dictionnaires de dictionnaires en dictionnaire de listes, j'ai eu un problème. J'avais extrait les données « dans l'ordre » d'apparition. Mais j'ai remarqué que dans chaque fichier, le premier event ne commence pas toujours par l'eid 'e1' mais peut commencer par 'e24' (deuxième fichier du dossier AQUAINT). Ce qui fait que lorsque les données sont prises dans l'ordre les events ne sont plus correctement reliés avec leur makeinstance :
1er event = e24
1er makeinstance = ei1 (devrait être ei24)

C'est pour cette raison que j'ai décidé de conserver ma structure en dictionnaires de dictionnaires pour créer mes deux CSV (event et timex).

En revanche, pour les identifiants artificiels, un simple dataframe en dictionnaire de listes fait l'affaire.

Ponctuation

Au cours des tests que j'ai pu faire sur Python j'ai remarqué que les signes de ponctuation étaient comptés comme des mots lors de l'étape de tokenisation. J'avais pris la décision de retirer cette ponctuation pour n'avoir que les mots. En discutant avec Anne-Lyse, nous avons conclu que la ponctuation pourrait être un indice temporel (par exemple les guillemets pour du discours rapporté). La ponctuation est actuellement conservée, nous pourrons en discuter lors de la réunion.

Contexte

Lors de l'extraction du contexte (event -4 / event +4) je me suis demandée si lorsque l'on a un event en fin de phrase, est-ce que l'on continue à prendre les mots +4 après l'event dans la phrase suivante ?

Actuellement, si on a un event en fin de phrase, le contexte event +4 sera vide (ou juste le point de fin de phrase).

Exemple :

Suspected bombs <EVENT eid="e1" class="OCCURRENCE">exploded</EVENT> outside the U.S. embassies in the Kenyan and Tanzanian capitals <TIMEX3 tid="t1" type="DATE" value="1998-08-07" temporalFunction="false" functionInDocument="NONE">Friday</TIMEX3>, <EVENT eid="e2" class="OCCURRENCE">killing</EVENT> dozens ['of', 'people', ',', 'witnesses'] **<EVENT eid="e3" class="REPORTING">said</EVENT>**.

['The', 'American', 'ambassador', 'to'] Kenya was among hundreds <EVENT eid="e12" class="OCCURRENCE">injured</EVENT>, a local TV <EVENT eid="e4" class="REPORTING">said</EVENT>.

Ou comme ce qui est fait actuellement dans mon script :

Suspected bombs <EVENT eid="e1" class="OCCURRENCE">exploded</EVENT> outside the U.S. embassies in the Kenyan and Tanzanian capitals <TIMEX3 tid="t1" type="DATE" value="1998-08-07" temporalFunction="false" functionInDocument="NONE">Friday</TIMEX3>, <EVENT eid="e2" class="OCCURRENCE">killing</EVENT> dozens ['of', 'people', ',', 'witnesses'] **<EVENT eid="e3" class="REPORTING">said</EVENT>** ['.']

The American ambassador to Kenya was among hundreds <EVENT eid="e12" class="OCCURRENCE">injured</EVENT>, a local TV <EVENT eid="e4" class="REPORTING">said</EVENT>.