Text Mining
&
Natural Language Processing

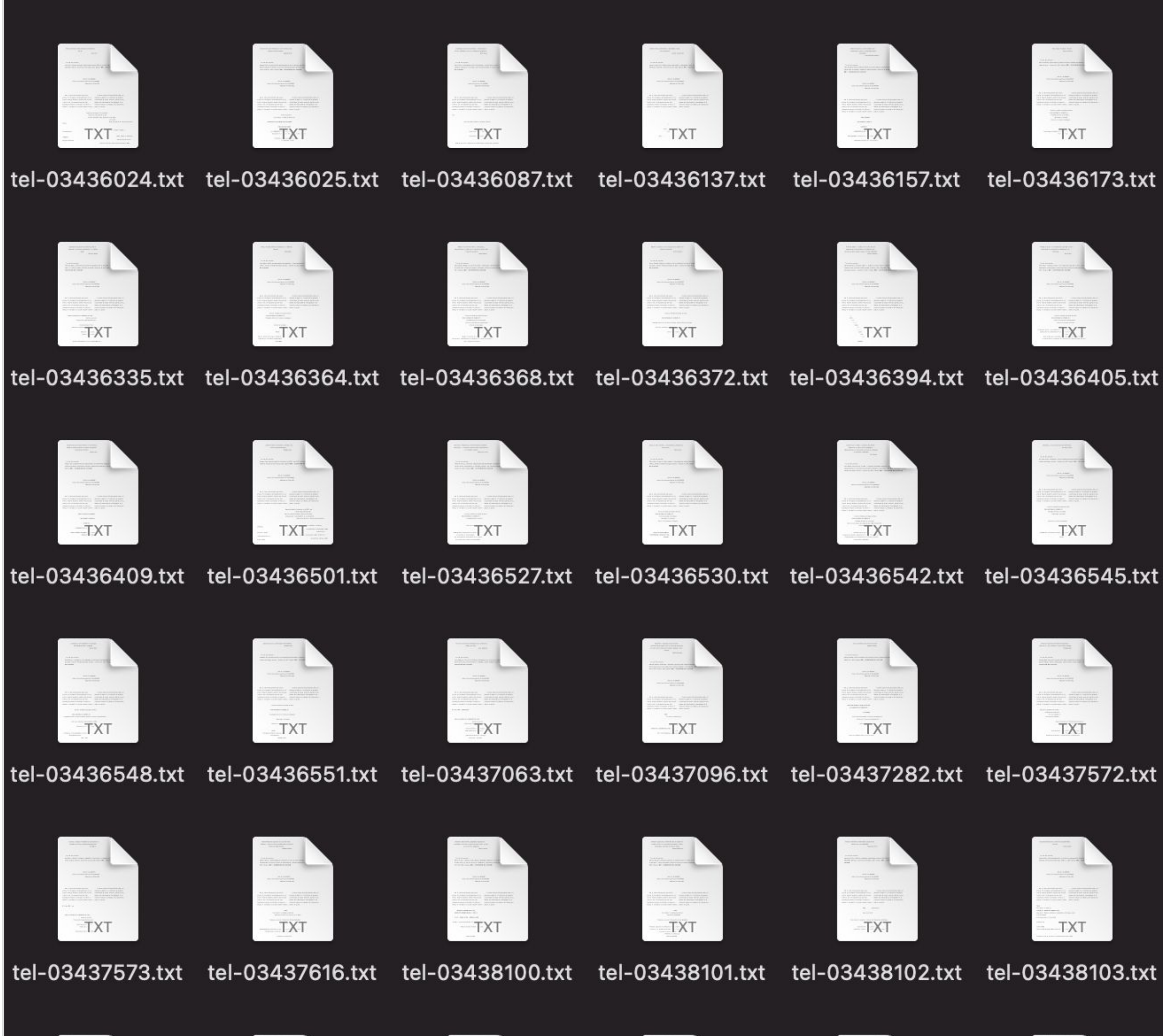
Anh Thu DOAN - PIB3

Table of Contents

1. Text Mining & doctoral dissertations
   1.1. Associated metrics (TF-IDF)
   1.2. Cosine similarity
2. POS-tagging and exercise generation
   2.1. Part-of-Speech Tagging
   2.2. Exercise generation

Text Mining & Doctoral dissertations

- Data used: A folder with 50 txt files of doctoral dissertations



tel-03436024.txt tel-03436025.txt tel-03436087.txt tel-03436137.txt tel-03436157.txt tel-03436173.txt

tel-03436335.txt tel-03436364.txt tel-03436368.txt tel-03436372.txt tel-03436394.txt tel-03436405.txt

tel-03436409.txt tel-03436501.txt tel-03436527.txt tel-03436530.txt tel-03436542.txt tel-03436545.txt

tel-03436548.txt tel-03436551.txt tel-03437063.txt tel-03437096.txt tel-03437282.txt tel-03437572.txt

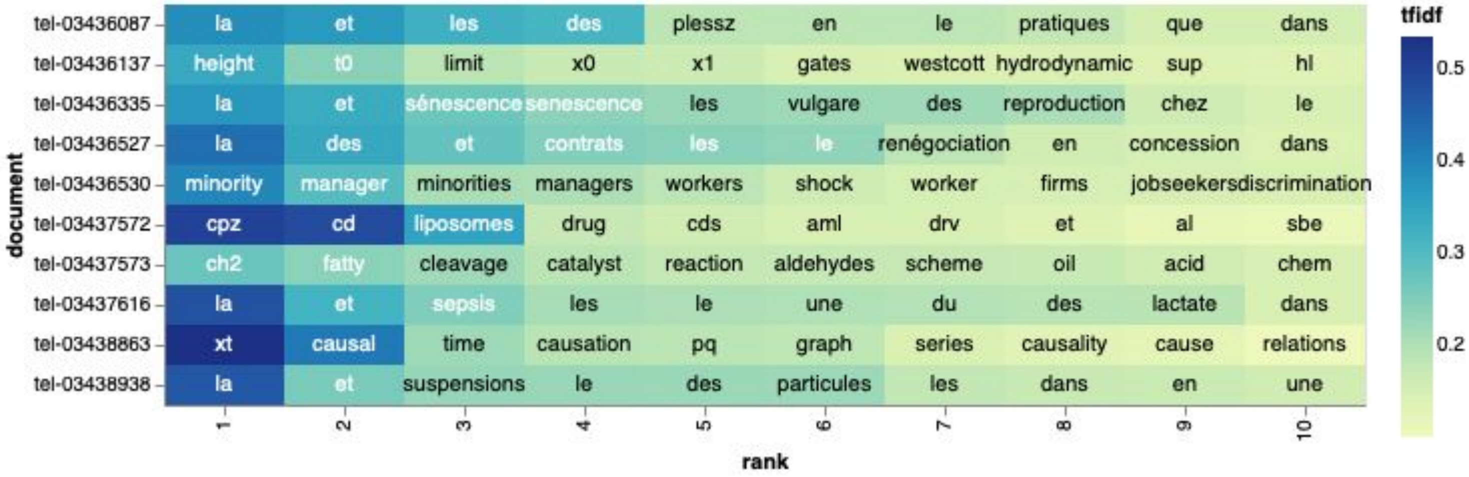
tel-03437573.txt tel-03437616.txt tel-03438100.txt tel-03438101.txt tel-03438102.txt tel-03438103.txt

Associated metrics (TF-IDF)

- *Tf-idf is a method for identifying the most frequently occurring or noteworthy terms in a document.*

- **Tf-idf** = **term_frequency** * **inverse_document_frequency**
  - **term_frequency** = *number of times a given term appears in document*
  - **inverse_document_frequency** = *log(total number of documents / number of documents with term) + 1******

Visualize TF-IDF



document	1	2	3	4	5	6	7	8	9	10
tel-03436087	la	et	les	des	plessz	en	le	pratiques	que	dans
tel-03436137	height	t0	limit	x0	x1	gates	westcott	hydrodynamic	sup	hl
tel-03436335	la	et	sénescence	senescence	les	vulgare	des	reproduction	chez	le
tel-03436527	la	des	et	contrats	les	le	renégociation	en	concession	dans
tel-03436530	minority	manager	minorities	managers	workers	shock	worker	firms	jobseekers	discrimination
tel-03437572	cpz	cd	liposomes	drug	cds	aml	drv	et	al	sbe
tel-03437573	ch2	fatty	cleavage	catalyst	reaction	aldehydes	scheme	oil	acid	chem
tel-03437616	la	et	sepsis	les	le	une	du	des	lactate	dans
tel-03438863	xt	causal	time	causation	pq	graph	series	causality	cause	relations
tel-03438938	la	et	suspensions	le	des	particules	les	dans	en	une

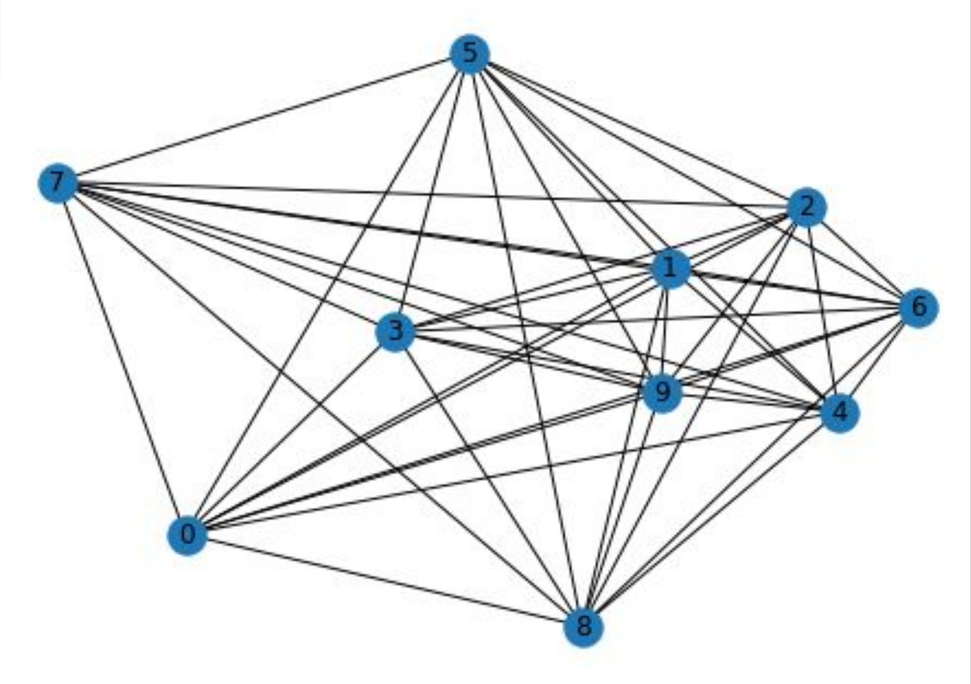
The highest TF-IDF scoring words of the first ten txt files

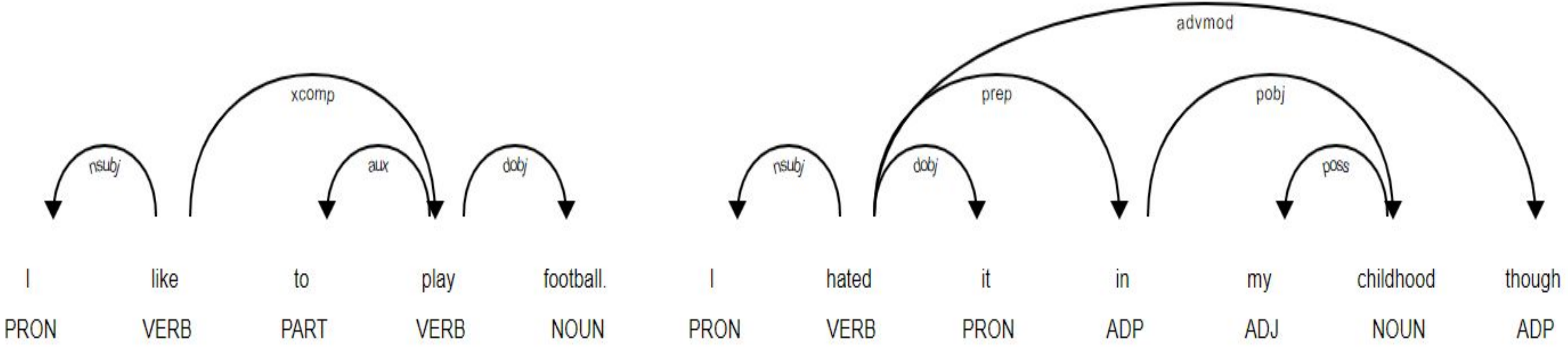
$$similarity(A,B) = \frac{A \cdot B}{\|A\| \text{ x } \|B\|} = \frac{\sum_{i=1}^{n} A_i \text{ x } B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \text{ x } \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Cosine similarity

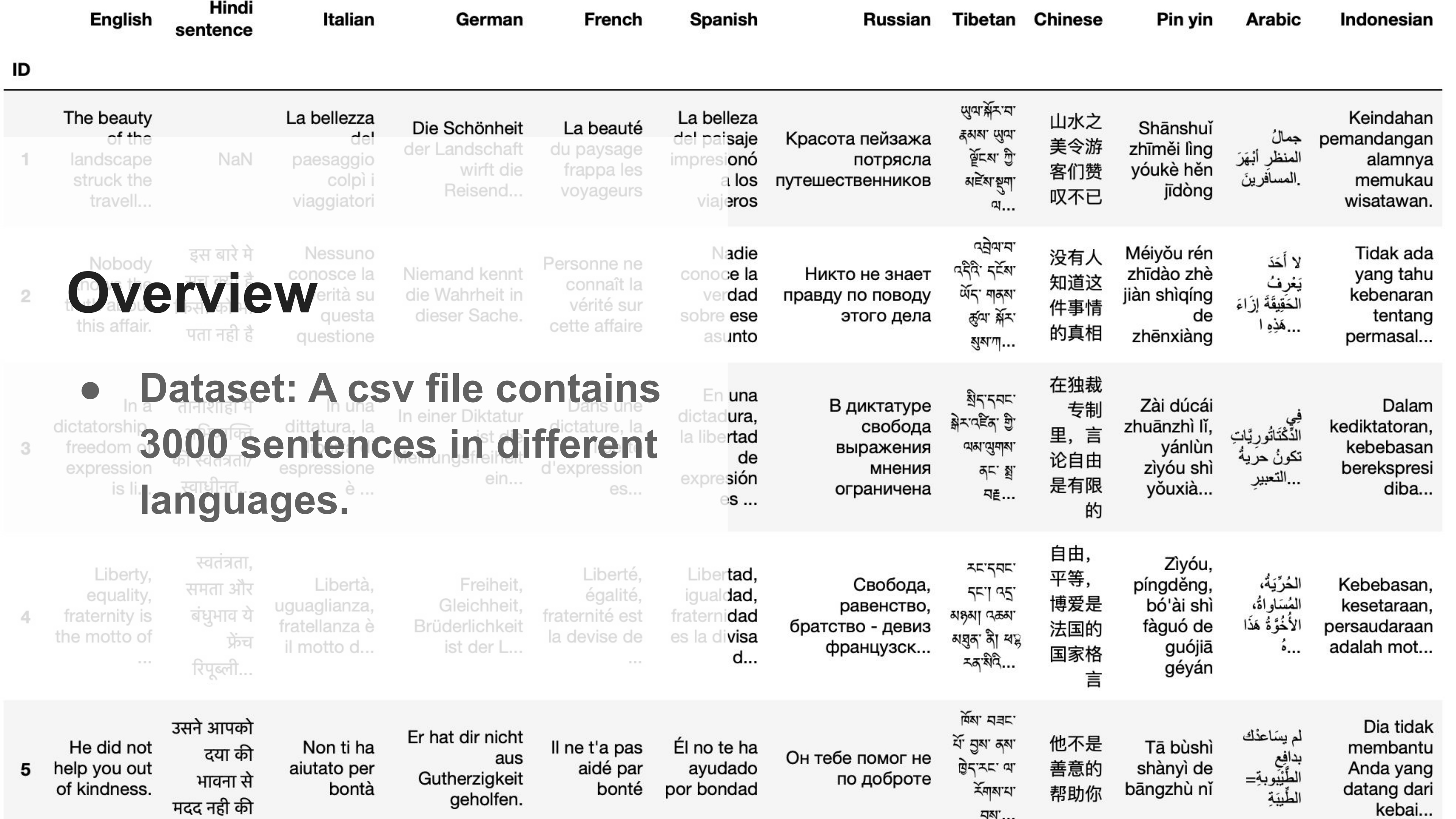
- The cosine similarity method calculates the cosine angle between two vector lists to determine how similar they are.
- In NLP, Cosine similarity is a metric for comparing the text similarity of two documents, regardless of their size.

Cosine similarity visualize for the first ten text file





POS-tagging & Exercise generation



Overview

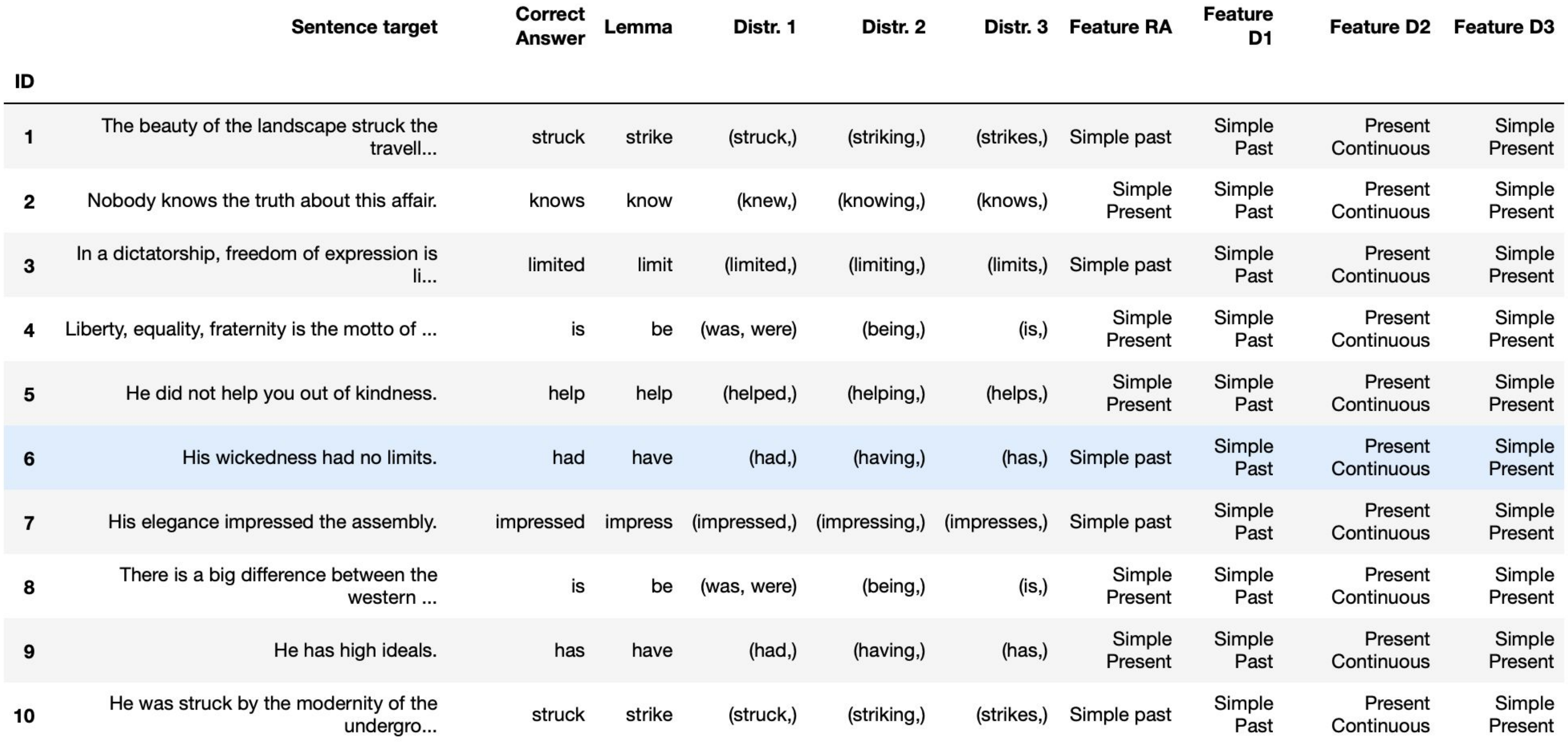
- **Dataset: A csv file contains 3000 sentences in different languages.**

	English	Hindi sentence	Italian	German	French	Spanish	Russian	Tibetan	Chinese	Pin yin	Arabic	Indonesian
ID												
1	The beauty of the landscape struck the travell...	NaN	La bellezza del paesaggio colpì i viaggiatori	Die Schönheit der Landschaft wirft die Reisend...	La beauté du paysage frappa les voyageurs	La belleza del paisaje impresionó a los viajeros	Красота пейзажа потрясла путешественников	ཡུལ་སྐོར་བ་ རྣམས་ ཡུལ་ ལྗོངས་ ཀྱི་ མཛེས་སྡུག་ ལ...	山水之美令游客们赞叹不已	Shānshuǐ zhīměi lìng yóukè hěn jīdòng	جمالُ المنظرِ أبْهَرَ المسافرينَ.	Keindahan pemandangan alamnya memukau wisatawan.
2	Nobody knows the truth about this affair.	इस बारे मे सच क्या है किसी को भी पता नही है	Nessuno conosce la verità su questa questione	Niemand kennt die Wahrheit in dieser Sache.	Personne ne connaît la vérité sur cette affaire	Nadie conoce la verdad sobre ese asunto	Никто не знает правду по поводу этого дела	འབྲེལ་བ་ འདིའི་ དངོས་ ཡོད་ གནས་ ཚུལ་ སྐོར་ སུས་ཀ...	没有人知道这件事情的真相	Méiyǒu rén zhīdào zhè jiàn shìqíng de zhēnxiàng	لا أَحَدَ يَعْرِفُ الحَقِيقَةَ إزَاءَ هَذِهِ ا...	Tidak ada yang tahu kebenaran tentang permasal...
3	In a dictatorship freedom of expression is li...	तानाशाही मे अभिव्यक्ति की स्वतंत्रता/ स्वाधीनत...	In una dittatura, la libertà di espressione è ...	In einer Diktatur ist die Meinungsfreiheit ein...	Dans une dictature, la liberté d'expression es...	En una dictadura, la libertad de expresión es ...	В диктатуре свобода выражения мнения ограничена	སྲིད་དབང་ སྐེར་འཛིན་ གྱི་ ལམ་ལུགས་ ནང་ སྨྲ་ བརྗ...	在独裁专制里，言论自由是有限的	Zài dúcái zhuānzhì lǐ, yánlùn zìyóu shì yǒuxià...	فِي الدِّكْتَاتُورِيَّاتِ تكونُ حريةُ التعبيرِ...	Dalam kediktatoran, kebebasan berekspresi diba...
4	Liberty, equality, fraternity is the motto of ...	स्वतंत्रता, समता और बंधुभाव ये फ्रेंच रिपूब्ली...	Libertà, uguaglianza, fratellanza è il motto d...	Freiheit, Gleichheit, Brüderlichkeit ist der L...	Liberté, égalité, fraternité est la devise de ...	Libertad, igualdad, fraternidad es la divisa d...	Свобода, равенство, братство - девиз французск...	རང་དབང་ དང་། འདྲ་ མཉམ། འཆམ་ མཐུན་ ནི། པཱ་རན་སིའི...	自由，平等，博爱是法国的国家格言	Zìyóu, píngděng, bó'ài shì fàguó de guójiā géyán	الحُرِّيَةُ، المُسَاواةُ، الأُخُوَّةُ هَذَا هُ...	Kebebasan, kesetaraan, persaudaraan adalah mot...
5	He did not help you out of kindness.	उसने आपको दया की भावना से मदद नही की	Non ti ha aiutato per bontà	Er hat dir nicht aus Gutherzigkeit geholfen.	Il ne t'a pas aidé par bonté	Él no te ha ayudado por bondad	Он тебе помог не по доброте	ཁོས་ བཟང་ པོ་ བྱས་ ནས་ ཁྱེད་རང་ ལ་ རོགས་པ་ བྱས...	他不是善意的帮助你	Tā bùshì shànyì de bāngzhù nǐ	لم يسَاعدْك بدافِعِ الطِّيبةِ=الطِّيبَةِ	Dia tidak membantu Anda yang datang dari kebai...

Methodology

- Two python library were used: spaCy, stanza
  - To process a particular language, we need to have Stanza language model as basic
  - In this part, we used some of feature of spaCy library such as: **Lemmatization, Part-of-speech(POS) Tagging**

Results



	Sentence target	Correct Answer	Lemma	Distr. 1	Distr. 2	Distr. 3	Feature RA	Feature D1	Feature D2	Feature D3
ID										
1	The beauty of the landscape struck the travell...	struck	strike	(struck,)	(striking,)	(strikes,)	Simple past	Simple Past	Present Continuous	Simple Present
2	Nobody knows the truth about this affair.	knows	know	(knew,)	(knowing,)	(knows,)	Simple Present	Simple Past	Present Continuous	Simple Present
3	In a dictatorship, freedom of expression is li...	limited	limit	(limited,)	(limiting,)	(limits,)	Simple past	Simple Past	Present Continuous	Simple Present
4	Liberty, equality, fraternity is the motto of ...	is	be	(was, were)	(being,)	(is,)	Simple Present	Simple Past	Present Continuous	Simple Present
5	He did not help you out of kindness.	help	help	(helped,)	(helping,)	(helps,)	Simple Present	Simple Past	Present Continuous	Simple Present
6	His wickedness had no limits.	had	have	(had,)	(having,)	(has,)	Simple past	Simple Past	Present Continuous	Simple Present
7	His elegance impressed the assembly.	impressed	impress	(impressed,)	(impressing,)	(impresses,)	Simple past	Simple Past	Present Continuous	Simple Present
8	There is a big difference between the western ...	is	be	(was, were)	(being,)	(is,)	Simple Present	Simple Past	Present Continuous	Simple Present
9	He has high ideals.	has	have	(had,)	(having,)	(has,)	Simple Present	Simple Past	Present Continuous	Simple Present
10	He was struck by the modernity of the undergro...	struck	strike	(struck,)	(striking,)	(strikes,)	Simple past	Simple Past	Present Continuous	Simple Present

Exercise generation

Q: The beauty of the landscape ___ the travellers.

A. Struck

B. Striking

C. Strikes

-> Correct Answer: A