Introduction –

For the CW0 (course work zero), our team, **The Novices** decided to choose the language French, of which the corpuses would be collected. Each member was asked to choose a country where French is a popular language even if not the first language and is spoken by a considerable proportion of the population of that country. The motive behind this idea was to create corpuses rich in French vocabulary and diverse enough to contain French words of each country's respective dialect of the language. I chose to collect the French corpus from Canada whereas the rest of the team focused on the countries – Ivory Coast, Morocco, Belgium, Switzerland and France.

Exploring & Experimenting on Sketch Engine-

At the beginning, I made progress by exploring the web interface of the Sketch Engine. Taking guidance from the article - Kilgarriff2014_Article_TheSketchEngineTenYearsOn.pdf , I began to understand and explore the different features available on Sketch Engine like Word Sketch, Concordance, Parallel Concordance, Thesaurus, Work Sketch Difference.

I experimented by generating word sketches to understand the grammatical usage of random words in English and that of the same words in different languages like French, Italian, Korean as well as in Hindi (language of my home country India) in their respective corpuses inbuilt in Sketch Engine. Although the word sketches generated for languages other than English and Hindi made little sense to me, but it led me to another feature in Sketch Engine which is Concordance. The concordance feature let me observe how difference words were used in context to a given word or how two or more given words were used in conjunction with one another and other words. For example, the French word for "Mountain" which is "Montagne" when used to generate word sketch of, showed up other French words which were used as verbs, modifiers, pre-positions etc. Using these words along with "Montagne" to check their concordance, let me observe how these verbs, modifiers, prepositions where being used in context with the word "Montagne". The parallel concordance feature helped put more sense into this as it helped me observe the concordance of words in a parallel corpora - for e.g. English and French and in my understanding these corpuses where exact translation of each other. The Thesaurus feature helped me discover synonyms and collocates.

Planning-

The team sat together to device a high-level plan to go about with the individual corpus collection. After some quick brainstorming an umbrella topic was fixed, and it was decided that the seed words the team would use would be pertinent to the topic as much as possible. The topic decided was **"Tourism"**. It was further decided that two of the seed words used by every team member would be common across the team and it was encouraged that the other seed words every team member used would be specific to that country's dialect as much as possible. The idea behind doing this was to ensure corpuses have general French words as well as French words specific to those dialects.

Collecting the Corpus-

I collected the corpus using the **New Corpus** feature in Sketch Engine in the page **MANAGE CORPUS** using the following steps –

- In the first step added a name for the corpus as **CW0_French_Canada**, selected the corpus type to be **Single Language Corpus** and set the language as **French**.
- In the next step for **ADD TEXTS**, selected "**Find texts on the web**" to perform a web search.
- In the third step, set further parameters for the collection – **Input type** as **Web search** and added the seed words – "**Tourisme**" meaning Tourism, "**Voyage**" meaning Travel, "**magasinage**" meaning shopping, "**Facture**" meaning bill.
- In the web search settings, switched on **Set values manually** and increased **Seed word search to 4** while keeping the value of **Max URLs Per Search to 30** and added the **TLD for Canada - .ca** under sites list.
- Next **pressed go** which led to another page with the list of web pages to download. I went through the list to ensure that they all have the TLD as **.ca**
- After this **hit go** again and let the corpus collection process kick off.

Since the collected corpus on several occasions reached very high words counts of the range of 150K to 200K, I tried to limit the word count by putting a limit on the parameter "**Max document size**" under web search settings however that was not of much help as the word counts were going to erratic numbers different times. Hence in the final attempt, when the corpus collection process reached ~70K, I manually cancelled the process and ensured that the collected corpus has compiled successfully by thoroughly monitoring it until a message on the screen showed successfully complied. Total words collected in the corpus is 70142.

Post collection, I checked the corpus by generating word sketch of the seed words, checking the concordance of the seed words.

Corpus comparison-

Another activity that the team performed together is corpus comparison of all the corpuses after the collections were done by the team members, using the compare corpora feature on Sketch Engine. The output of the same is the adjacent matrix where 1 is a complete match and differences increases as the score goes higher.

	CH	BE	CA	FR	CI	MA
FrenchCH	1.00	4.40	5.06	5.59	4.01	4.48
FrenchBE	4.40	1.00	3.64	3.61	4.06	4.11
FrenchCA	5.06	3.64	1.00	4.15	4.75	4.64
FrenchFR	5.59	3.61	4.15	1.00	5.40	5.45
FrenchCI	4.01	4.06	4.75	5.40	1.00	3.72
FrenchMA	4.48	4.11	4.64	5.45	3.72	1.00