République Algérienne Démocratique & Populaire
Ministère de l'Enseignement Supérieur & de la Recherche Scientifique
Ecole Supérieure en Informatique 08 mai 1945 de Sidi Bel Abbes
Année universitaire : 2025-2026 **Semestre 1**
3 SC : SIW & ISI **Module : Big Data**

TP 3 : Spark SQL

Sous Google Colab, installer pyspark :

```
! pip install pyspark
```

Prenons un exemple où nous créons et utilisons un DataFrame. Nous utiliserons le dataset N-Gram de Google[1]. Dans cet exemple. Le fichier **ngram.csv** contient des données sur les bigrammes avec les colonnes suivantes : ngram : String, Year : int, Count : int, Pages : int, Books : int.

Chaque ligne contient :

- **Bigram :** le bi-gramme lui-même.
- **Year :** année d'apparition du bi-gramme.
- **Count :** nombre de fois que le bi-gramme est apparu dans les livres de l'année correspondante.
- **Pages :** nombre de pages sur lesquelles le bi-gramme est apparu dans l'année Year (page-count).
- **Books :** nombre de livres distincts dans lesquels le bi-gramme est apparu dans l'année Year (book-count).

1) Créer un Dataframe à partir du fichier **ngram.csv.**
2) Enregistrer le Dataframe créé en tant que table temporaire.
3) Répondre aux requêtes suivantes en utilisant deux méthodes : (1) le langage SQL, et (2) les méthodes de l'API SparkSQL.
   1) Retourner tous les bi-grammes dont le nombre Count est supérieur à cinq.
   2) Retourner le nombre total de bi-grammes dans chaque année.
   3) Retourner les bi-grammes qui ont le plus grand nombre de count dans chaque année.
   4) Retourner tous les bi-grammes qui sont apparus dans 20 années différentes.
   5) Retourner tous les bi-grammes qui contiennent le caractère '!' dans la première partie et le caractère '9' dans la deuxième partie (les deux parties sont séparées par un espace).
   6) Retourner les bi-grammes qui sont apparus dans toutes les années présentes dans les données.
   7) Retourner le nombre total de pages et de livres dans lesquels chaque bi-gramme apparaît pour chaque année disponible, trié par ordre alphabétique.

[1] Google NGram Dataset : http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

3.8) Retourner le nombre total de bi-grammes différents dans chaque année, triés par ordre décroissant de l'année.