MASTER EN SCIENCES ET TECHNIQUES

SDI

SYSTÈMES D'INFORMATION DÉCISIONNELS ET IMAGERIE

FACULTÉ DES SCIENCES ET TECHNIQUES ERRACHIDIA

جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

كلية العلوم والتقنيات
FACULTÉ DES SCIENCES ET TECHNIQUES

Classification des articles utilisant mahout

Présenté par

Mohamed Harafi

Mohamed El Azzouzi

Encadré par

Pr. Bakhoya Mohamed

PLAN

1. Introduction et problémtique
2. Avantages d’utilisation de Hadoop en ML
3. Prétraitement de données
4. Extraction de caractéristiques
5. Modélisation et entraînement
6. Conclusion

Introduction

La classification de texte est une tâche essentielle en traitement automatique des langues (TAL), visant à attribuer une étiquette ou une catégorie à un texte donné. Cette tâche trouve des applications dans divers domaines tels que :

- La détection de spams.
- L'analyse des sentiments.
- La catégorisation de documents.

Comment Mahout sur Hadoop permet-il une classification de textes efficace, rapide et évolutive sur de grands volumes de données ?

Hadoop en ML

Définition

> Hadoop est une plateforme open-source conçue pour le stockage et le traitement distribué de grandes quantités de données.
> Elle utilise le système de fichiers distribué Hadoop (HDFS) pour le stockage et le modèle de programmation MapReduce pour le traitement des données.

Hadoop en ML

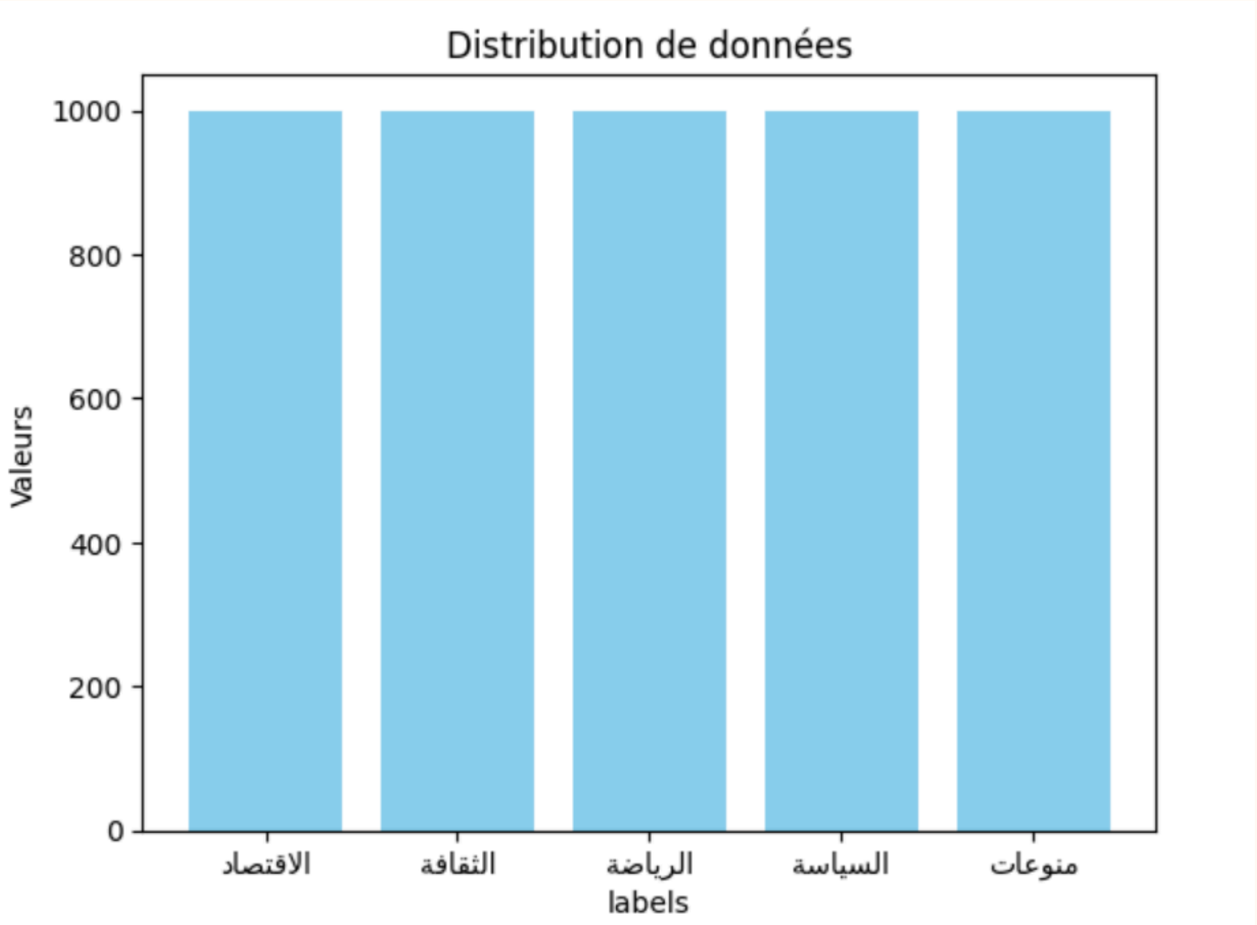
Avantages

- Gestion efficace de vastes ensembles de données nécessaires à l'entraînement des modèles de machine learning.
- Traitement parallèle des données, accélérant ainsi les processus d'entraînement et d'évaluation des modèles.
- Intégration avec des bibliothèques de machine learning telles qu'Apache Mahout et Spark MLlib pour des algorithmes évolutifs.

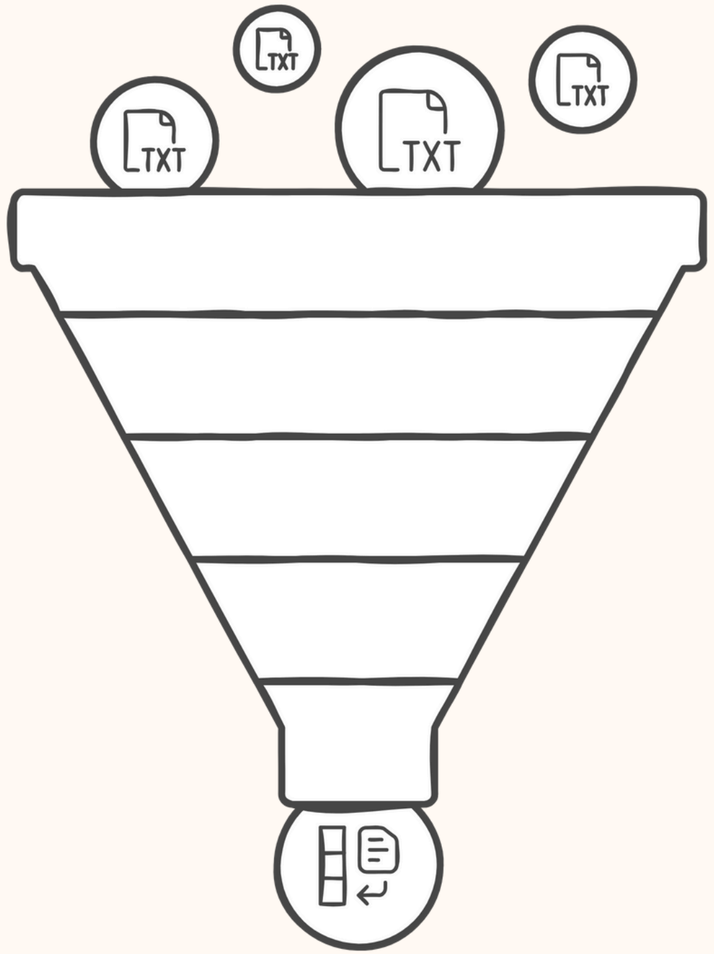
Exploration du Dataset

	text	label
0	جب دسم جمع كثير كوميدي بعض طرب مفاجات ثير خري ...	0
1	سجل ممثل مغرب سعيد غماو حضور ضمن فيلم سينماء ج...	0
2	ستعد مطرب نسيم محمد اطلاق عمل غناء جديد عنوان ...	0
3	اط قيصر غناء عرب نجم عراق اظم ساهر عبر اذاع مي...	0
4	عبدالال وسحاب خبار مغرب اطار احتفال شعب مغرب ع...	0
...	...	...
4995	ضت قرع دور مجموع اقصاء افريق مءهل نسخ قادم اس ...	4
4996	حيد معد دنيا نبخت صغير واصل دريب حراس مرمي جري...	4
4997	وم ظهر صول جديد مسلسل ضيح مبار ره قدم مغشوش سا...	4
4998	مفتال منضم مولود وجد قال ستفاد تجرب كوديم قال ...	4
4999	رابح حشلاف دخل جواء طول عالم زين ختبر حظوظ الا...	4

5000 rows × 2 columns



Prétraitement de données



Supprimer les mots vides

Appliquer le stemming léger

Normaliser le texte arabe

Supprimer les éléments non textuels

Organiser par catégorie

Structurer les documents par thème

Extraction des caractérisques

Pour L'extraction de caractéristiques depuis les données textuelles nous avons utiliser TF-IDF (Term Frequency-Inverse Document Frequency), qui consiste à transformer un ensemble de textes (documents) en une représentation numérique.

```
:2.7170231342315674,7175:3.8404393196105957,5159:3.972015857696533,11888:3.2926347255706787,11953:3.2711379528045654,42
39:3.9997403621673584,2889:3.7116763591766357,2984:8.173480033874512,5492:4.857787132263184,10373:3.128631830215454,134
76:3.764331579208374,5328:6.572754383087158,3150:7.096971035003662,9024:5.645992279052734,11294:3.577021837234497,15645
:3.7870934009552,5298:3.4057259559631348,6179:4.863232612609863,15291:6.299752712249756,8826:3.9004220962524414,473:2.9
58995342254638 7,7737:3.4281482696533203,7071:5.062845706939697,10304:3.7120583057403564,2525:6.381699085235596,7575:3.8
6470389366149 9,13611:5.509860038757324,15127:4.5027337074279785,3015:13.113846778869629,15613:4.48675537109375,12249:4.
092243194580078,12251:3.476938486099243,15619:6.1714887619018555,8652:4.174859046936035,86:4.381394863128662,14694:3.21
64072990417 48,5862:4.3242363929748535,10480:2.8275938034057617,13416:3.1928341388702393,15642:4.2860846519470215,6109:2
.229632616043091,6570:3.419118881225586,6157:4.670560359954834,3733:6.02068567276001,122:4.5684332847595215,6854:4.6119
184494018555,10363:3.1353771686553955,1887:6.586530685424805,4064:6.932578086853027,1296:3.6118314266204834,12129:7.907
755374908447,2137:7.1274333000183105,9950:6.11599588394165,8943:1.9127968549728394,3360:5.4063191413879395,1769:2.54834
2704772949,13064:4.25455379486084,870:3.861201286315918,15687:2.718137264251709,2757:7.509149551391602,9339:3.234926462
173462,11995:4.863232612609863,10173:5.774900436401367,12436:5.906275272369385,5789:3.4557361602783203,13760:5.91092491
1499023,9821:3.3005871772766113,10153:10.979711532592773,4789:3.9542996883392334,15703:4.249334812164307,13853:2.932401
8955230713}
Count: 5000
25/02/01 11:59:14 INFO MahoutDriver: Program took 6218 ms (Minutes: 0.10363333333333333)
hadoop@harafi:~$
```

Modélisation

Naïve Bayes

Le Naive Bayes est une méthode très populaire et efficace pour la classification de texte, notamment en raison de sa simplicité, de sa rapidité et de ses bonnes performances dans de nombreux cas de classification de texte.

Statistics	
Kappa	0,9564
Accuracy	97,1353%
Reliability	80,9476%
Reliability (standard deviation)	0,3968
Weighted precision	0,9716
Weighted recall	0,9714
Weighted F1 score	0,9714

Modélisation

Naïve Bayes

```
Summary
-------------------------------------------------------
Correctly Classified Instances          :       2950       97,1353%
Incorrectly Classified Instances        :         87        2,8647%
Total Classified Instances              :       3037

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       <--Classified as
575     0       0       24      4        |  603          a     = الاقتصاد
1       577     1       3       7        |  589          b     = الثقافة
0       0       603     3       3        |  609          c     = الرياضة
14      1       1       587     11       |  614          d     = السياسة
4       0       1       9       608      |  622          e     = منوعات
```

Modélisation

Naïve Bayes : Résultats de classificaion sur les données de test

Statistics	
Kappa	0,878
Accuracy	91,2379%
Reliability	76,0814%
Reliability (standard deviation)	0,3749
Weighted precision	0,9147
Weighted recall	0,9124
Weighted F1 score	0,9128

Modélisation

Naïve Bayes : Résultats de classificaion sur les données de test

```
Summary
-------------------------------------------------------
Correctly Classified Instances          :       1791       91,2379%
Incorrectly Classified Instances        :        172        8,7621%
Total Classified Instances              :       1963

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       <--Classified as
348     4       0       37      8        |  397          a     = الاقصاد
10      367     3       20      11       |  411          b     = الثقافة
4       6       372     5       4        |  391          c     = الرياضة
22      5       2       337     20       |  386          d     = السياسة
2       1       0       8       367      |  378          e     = منوعات
```

Conclusion

l'utilisation de Hadoop pour la classification de textes est une solution puissante pour traiter de grands volumes de données de manière distribuée et évolutive. Bien que cela permette d'améliorer la scalabilité et de réduire les coûts, la mise en œuvre peut être complexe et nécessiter des ajustements pour optimiser les performances et le traitement des données. Pour des projets à grande échelle, Hadoop est un excellent choix, mais il demande une bonne maîtrise technique et une gestion adéquate des ressources.

Merci pour votre attention