POSTER PROJET LU3IN026

Dans ce projet, nous avons étudié les performances de plusieurs algorithmes d'apprentissage automatique sur les données textuelles du forum 20 Newsgroups, qui contient des messages issus de 20 forums de discussion différents.
Nous avons appliqué à la fois des approches supervisées en approche supervisée et non-supervisées (clustering), en expérimentant différentes représentations vectorielles du texte : bag-of-words, fréquences, TF-IDF, avec ou sans suppression des stopwords.
L'objectif est d'appliquer et comparer différents algorithmes d'apprentissage supervisé et non-supervisé sur des données textuelles. Étudier l'influence des représentations et des classifieurs sur la performance (accuracy, temps, robustesse).

ECHANTILLON

	messages	target
0	\n\nI am sure some bashers of Pens fans are pr...	10
1	My brother is in the market for a high-perform...	3
2	\n\n\n\n\tFinally you said what you dream abou...	17
3	\nThink!\n\nIt's the SCSI card doing the DMA t...	3
4	1) I have an old Jasmine drive which I cann...	4
...	...	...
18841	DN> From: nyeda@cnsvax.uwec.edu (David Nye)\nD...	13
18842	\nNot in isolated ground recepticles (usually ...	12
18843	I just installed a DX2-66 CPU in a clone mothe...	3
18844	\nWouldn't this require a hyper-sphere. In 3-...	1
18845	After a tip from Gary Crum (crum@fcom.cc.utah....	7

API

- Python
- NumPy
- Pandas
- Graphviz
- Matplotlib

SCIENCES SORBONNE UNIVERSITÉ

GUILLAUME DUPART
YANIS TOUTAIN
GROUPE 2
LU3IN026 - 2024 2025

PRÉTRAITEMENT

NETTOYAGE DU TEXTE

- Nettoyage → Met le texte en minuscule, remplace la ponctuation par un espace
- text2vect → Supprime les mots inutiles (conjonction, verbes trop communs ...).

REPRÉSENTATION VECTORIELLES

- Bag-of-words binaire (1 si au moins une fois dans le document, 0 sinon).
- Bag-of-words par comptage (compteur de mots par document)
- Fréquence relative (compteur mot / total)
- TF-IDF (l'importance d'un mot dans un document donné, tout en réduisant le poids des mots trop fréquents dans l'ensemble du corpus).

DONNÉES

- 18846 lignes
- Train (avec stop-words) : 173 lignes
- Test (avec stop-words) : 932 lignes
- Train (sans stop-words) : 173 lignes
- Test (sans stop-words) : 932 lignes

APPROCHE SUPERVISÉE

KNN AVEC DISTANCE EUCLIDIENNE

In this section, state what is the purpose of your study.

KNN AVEC DISTANCE COSINUS

In this section, state what is the purpose of your study.

PERCEPTRON

In this section, state what is the purpose of your study.

ARBRE DE DÉCISION

In this section, state what is the purpose of your study.

APPROCHE NON SUPERVISÉE

CONCLUSION

To wrap up your poster, present two to three key findings. You can also add a brief explanation or narrative to these that can encourage conversation or dialogue with the audience. These findings can be actionable items that can lead to implementation, policy creation or further study.

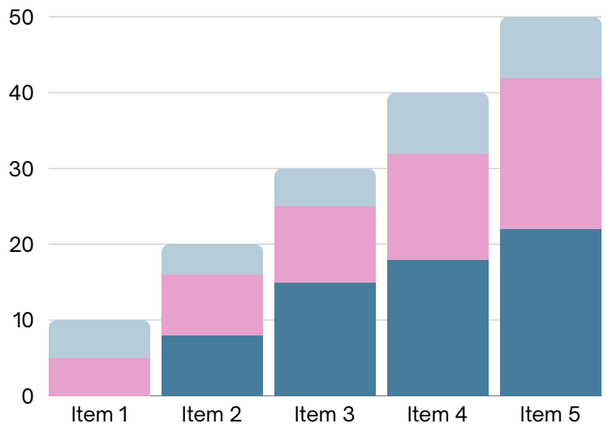
Related literature
References can take up a lot of space, so cite only the key references used in the study.

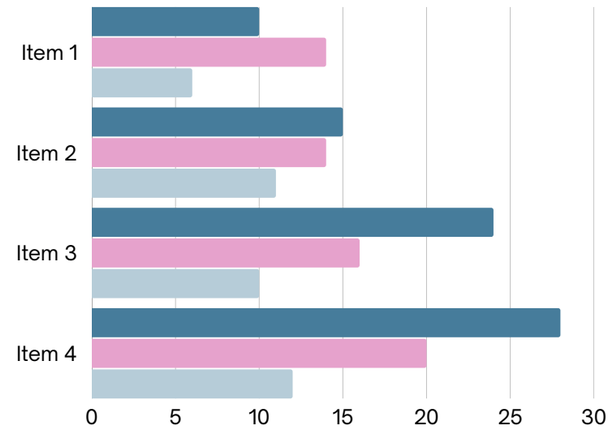
CHARTS

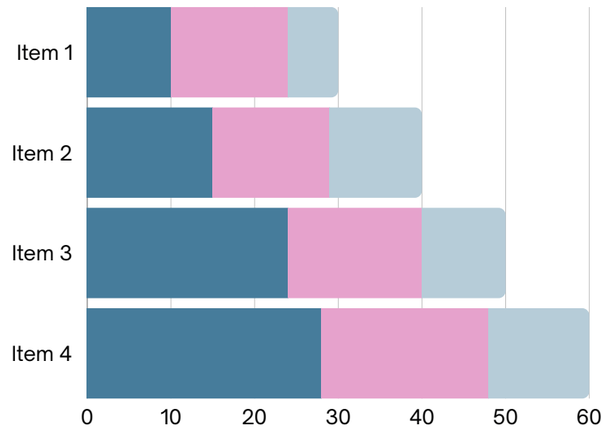
40
30
20
10
0
Item 1
Item 2
Item 3
Item 4
Item 5

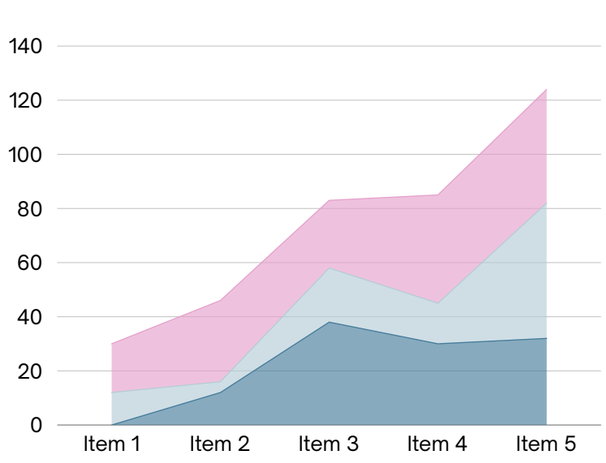
50
40
30
20
10
0
Item 1
Item 2
Item 3
Item 4
Item 5

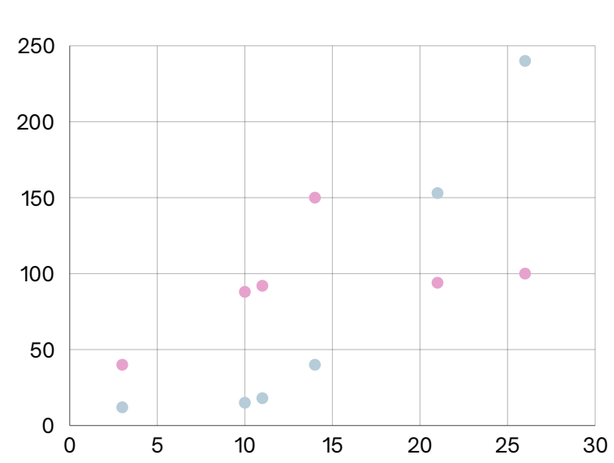
25
20
15
10
5
0
Item 1
Item 2
Item 3
Item 4
Item 5

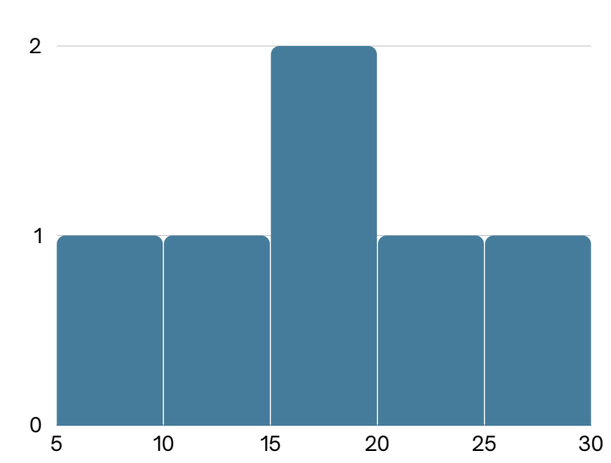

50
40
30
20
10
0
Item 1
Item 2
Item 3
Item 4
Item 5

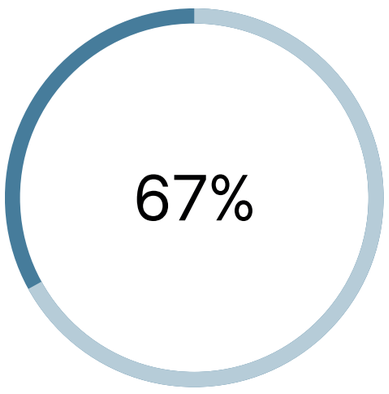

Item 1
Item 2
Item 3
Item 4
0
5
10
15
20
25
30

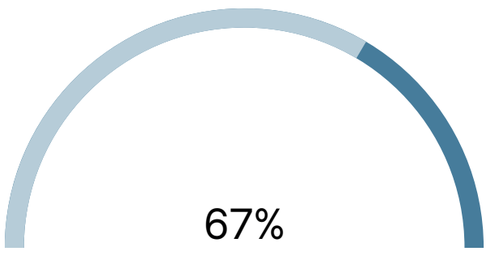

Item 1
Item 2
Item 3
Item 4
0
10
20
30
40
50
60


140
120
100
80
60
40
20
0
Item 1
Item 2
Item 3
Item 4
Item 5


250
200
150
100
50
0
0
5
10
15
20
25
30


2
1
0
5
10
15
20
25
30


67%


67%


67%