

# Présentation finale

## Vectorisation de texte

Hanane, Farouq, Walid, Asma, Steve

Université d'Avignon

6 janvier



# Plan

- 1 Démo finale
- 2 Projet et méthode Scrum
- 3 Bilan

# US2 – Fréquence des mots

- On colle un texte.
- On affiche les mots les plus fréquents.

← Retour

## Fréquence des mots

Texte :

mot s'applique à l'enchevêtrement des fibres utilisées dans le tissage, voir par exemple Ovide : « Quo super *infecti textum rute sedula baucis* = (un siège) sur lequel Baucis empressée avait jeté un tissu qu'il empêtrait[2] ou au tressage (exemple chez Martial = *Vinumque textor* = pavier d'osier tressé → vers 1er siècle ap. J.-C.) ou au large de construire comme dans « *baillique* » chez Ciceron[3]. »

Le sens figuré d'éléments de langage organisés et enchaînés apparaît avant l'empire romain : il désigne un agencement particulier du discours. Exemple : « *epistolas tenere* » composer des épîtres

Lancer l'analyse

Mot	Fréquence
siecle	6
mot	5
texte	4
chez	3
exemple	3
sen	3
textore	3
textum	3
agencement	2
applique	2
baucis	2
ciceron	2
comme	2
construire	2
ier	2

# US4 – Similarité entre deux textes

- On colle deux textes.
- On obtient un score de similarité.

[-- Retour](#)

## Similarité entre deux textes

### Texte 1 :

— Ciceron [1er siècle av. J.-C.]<sup>[4]</sup> ou plus nettement chez Quintilien [1er siècle apr. J.-C.] : « verba in textu jungantur = l'agencement des mots dans la phrase »<sup>[5]</sup>.

Les formes anciennes du Moyen Âge désignent au XIIe siècle le volume qui contient le texte sacré des Évangiles, puis au XIIIe siècle, le texte original d'un livre saint ou des propos de quelqu'un. Au XVIe siècle, le mot s'applique au passage d'un ouvrage pris comme référence et au début du XIXe siècle le mot texte a son sens général d'"écrit".

### Texte 2 :

de texte brut dans un protocole de télécommunication ;  
de document texte si le logiciel utilisé permet de формater le texte ;  
ou encore de texte formaté ou riche lorsque les indications de formats sont données en texte brut  
(exemple : Rich Text Format) ;  
de fichier texte dans un fichier qui ne contient que des caractères sans mise en forme autre que  
les sauts de lignes et le choix des caractères.

[Comparer](#)

Similarité : 3.13 %

# US6 – Snippets : meilleur passage

- On donne un document + une requête.
- On affiche le meilleur passage (fenêtre).

## Moteur de Recherche de Snippets

### 1. Document (Texte source)

Le sens figuré d'éléments de langage organisés et enchainés apparaît avant l'Empire romain : il désigne un agencement particulier du discours. Exemple : « *epistolas texere* = composer des épîtres » - Ciceron (Ier siècle av. J.-C.)[4] ou plus nettement chez Quintilien (Ier siècle apr. J.-C.) : « *verba in textu jungantur* = l'agencement des mots dans la phrase »[5].

Les formes anciennes du Moyen Âge désignent au XII<sup>e</sup> siècle le volume qui contient le texte sacré des Évangiles, puis au XIII<sup>e</sup> siècle, le texte original d'un livre saint ou des propos de quelqu'un. Au XVII<sup>e</sup> siècle le mot s'applique au passage d'un ouvrage pris comme référence et au début du XIX<sup>e</sup> siècle le mot texte à son sens général d'"écrit" »[6].

### 2. Requête (Mots clés)

texte

#### Taille Fenêtre (mots)

3

#### Nombre de phrases (Top K)

3++

Activer la pondération TF-IDF (si décoché : comptage simple)

 Meilleur Passage (Fenêtre)

 Top Phrases Clés

#### Meilleur Passage (Contexte continu)

Score: 0.7631 | TF-IDF | Position mot: 0

« Texte »

# US6 – Snippets : Top phrases (Top K)

- On choisit Top K.
- On affiche les phrases les plus pertinentes.

**Moteur de Recherche de Snippets**

**1. Document (Texte source)**

Le sens figuré d'éléments de langage organisés et enchaînés apparaît avec l'empire romain : il désigne un agencement particulier du discours. Exemple : « apothéose tessera » = composer des épitres = « Cidren [er] iuste et fortis] / quod est in libro de laudibus Quodlibetum [littera vobis agri, 3, 6, 1] » = morte et morte impetratur ». L'agencement des mots sera dans la phrase «[qui]».

Les forces anciennes du Moyen Âge désignent au XIIe siècle le volume qui contient le texte sacré des Évangiles, puis au XIIIe siècle, le texte original d'un livre saint ou des propositus de quelqu'un.

Les forces anciennes du Moyen Âge désignent au XIIe siècle le volume qui contient le texte sacré des Évangiles, puis au XIIIe siècle, le texte original d'un livre saint ou des propositus de quelqu'un.

Le sens figuré d'éléments de langage organisés et enchaînés apparaît avec l'empereur romain : il désigne le sens figuré au passage d'un ouvrage pris comme référence et au début du XIIe siècle le mot

**2. Requête (Mots clés)**

texte

**Taille Fenêtre (mots)**  **Nombre de phrases (Top K)**

Activer la pondération TF-IDF (si décoché : couplage simple)

Meilleur Passage (Fenêtre)  Top-Phrases Clés

**Top N Phrases Pertinentes**

#1 | Score 1.5281 | index phrase ?  
"Les formes anciennes du Moyen Âge désignent au XIIe siècle le volume qui contient le texte sacré des Évangiles, puis au XIIIe siècle, le texte original d'un livre saint ou des propositus de quelqu'un."

#2 | Score 0.7621 | index phrase ?  
"Texte" = est issu du mot latin "textum", dérivé du verbe "texere" qui signifie "tisser"."

#3 | Score 0.7621 | index phrase ?  
"Au XIIe siècle le mot s'applique au passage d'un ouvrage pris comme référence et au début du XIIe siècle le mot texte a son sens général d'"écrire" [1]."

#4 | Score 0 | index phrase ?  
"Le mot s'applique à l'enrelachement des fibres utilisées dans le tissage, voir par exemple Ovide : "Quo super interiect tessera ruda sedula Bauci" [un siège] sur lequel Bauci empressa arant jecu[n]dum tenuissim tessera [2] ou au tressage (exemple chez Martial et Viminenus tessera = parier d'osier tresser)"

#5 | Score 0 | index phrase ?  
"Le verbe a aussi le sens large de constituer comme dans "basilicam tessera" = construire une basilique = chez Cidren [3]."

#6 | Score 0 | index phrase ?  
"Le sens figuré d'éléments de langage organisés et enchaînés apparaît avec l'Empire romain : il désigne un agencement particulier du discours."

#7 | Score 0 | index phrase ?  
"Exemple : "apothéose tessera" = composer des épitres = "Cidren [er] iuste et fortis] /

# US5 – Résumé automatique

- On choisit un pourcentage à garder.
- On génère le résumé avec une méthode.

← Retour

## Résumé automatique

### Texte à résumer :

Le sens figuré d'éléments de langage organisés et enchaînés apparaît avant l'Empire romain : il désigne un agencement particulier du discours. Exemple : « epistolas texere = composer des épîtres » - Cicéron (Ier siècle av. J.-C.)[4] ou plus nettement chez Quintilien (Ier siècle apr. J.-C.) : « verba in textu jungantur = l'agencement des mots dans la phrase »[5].

Le sens figuré d'éléments de langage organisés et enchaînés apparaît avant l'Empire romain : il désigne un agencement particulier du discours. Exemple : « epistolas texere = composer des épîtres » - Cicéron (Ier siècle av. J.-C.)[4] ou plus nettement chez Quintilien (Ier siècle apr. J.-C.) : « verba in textu jungantur = l'agencement des mots dans la phrase »[5].

Les formes anciennes du Moyen Âge désignent au XIIe siècle le volume qui contient le texte sacré des Évangiles, puis au XIIIe siècle, le texte original d'un livre saint ou des propos de quelqu'un. Au XVIIe siècle le mot s'applique au passage d'un ouvrage pris comme référence et au début du XIXe siècle le mot texte a son sens général d'« écrit »[6].

Pourcentage à garder :  %

Méthode de résumé :

Premières phrases

Générer le résumé

« Texte » est issu du mot latin « *textum* », dérivé du verbe « *texere* » qui signifie « *tisser* ». Le mot s'applique à l'entrelacement des fibres utilisées dans le tissage, voir par exemple Ovide : « *Quo super iniecit textum rude sedula Baucis* = (un siège) sur lequel Baucis empressée avait jeté un tissu grossier »[2] ou au tressage (exemple chez Martial « *Vimineum textum = panier d'osier tressé* »).

# US7 – Ponctuation automatique

- On donne un texte sans ponctuation.
- On obtient une version plus lisible.

[← Retour au menu](#)

## Ponctuation automatique

Texte sans ponctuation :

il était une fois dans un pays lointain un roi sage et juste  
il gouvernait avec bonté son peuple était heureux et prospère  
un jour un dragon terrifiant apparut dans les montagnes il semait la terreur

[Ajouter la ponctuation](#)

## Résultat :

Il était une fois. Dans un pays loin. Tain. Un roi sage et juste, il gouvernait avec bon, té, son peuple était heureux et prosp, ère, un jour. Un drago n ter ri fiant apparut dans les montagne, s, il semait la terre, ur,.

# Sujet du projet

- Thème : application web de traitement de texte.
- On a développé : TF-IDF, similarité, résumé, snippets, ponctuation.
- Objectif : livrer petit à petit en appliquant Scrum.

# Product Backlog

- On a écrit nos besoins en **User Stories** : interface, TF-IDF, similarité, résumé, snippets, ponctuation, idée LLM.
- Pourquoi : pour **prioriser** ce qui est important.
- Exemple : l'US **LLM** est restée à *faire* car on a d'abord livré les modules de base.

D	User Story	Definition of Done (DoD)	Description / Fonctionnalité	Sprint	Statut
US1	En tant qu'utilisateur, je veux une <b>page d'accueil web claire</b> avec les 4 modules	- Page responsive- 4 boutons fonctionnels- Navigation fluide	Interface de navigation pour rentrer le texte à résumer + la taille , un bouton pour afficher un les mots les plus fréquents en haut dans le texte rentré ....	1,2,3,4	In progress
US2	En tant qu'utilisateur, je veux <b>afficher les vecteurs de mots</b> (mots les plus fréquents en haut)	- Upload texte- Calcul fréquence- Top 10 mots affichés	Visualisation statistique des mots importants	1	Fait
US3	En tant qu'utilisateur, je veux calculer <b>de TF et IDF</b>	- Saisie texte- Calcul TF/IDF- Affichage tableau	- Saisie texte- Calcul TF/IDF- Affichage tableau		Fait
US4	En tant qu'utilisateur, je veux comparer deux textes pour détecter une <b>similitude / plagiat</b>	- 2 zones texte- Score 0-1- Visualisation	Calcul de similarité entre textes	2,3	Pas commencé
US5	En tant qu'utilisateur, je veux coller un texte et obtenir un <b>résumé</b>	- Choix % réduction- Résumé cohérent- Mots clés conservés	Module de résumé automatique avec choix de la taille du résumé	2, 3	Pas commencé

# Sprints

- On a fait **6 sprints** au total.
- Un sprint = **une séance de TP** (sprint court).
- Pourquoi : en 3h, on vise un **résultat visible** (page, score, module utilisable).

# Sprint Backlog + Kanban

- Au début du sprint : on choisit les US du sprint.
- On les découpe en **tâches** et on les met sur un tableau **Kanban**.
- On attribue les tâches selon la difficulté et les capacités de chacun.

Ao taches	Date	hana	État
📄 rectification de TF-IDF et des stopwor	4 décembre 2025 16:00 → 17	H Hanan Boud	Terminé
📄 resumé les 3 méthodes	4 décembre 2025 16:00 → 17	A Asma Nihal Boukraa	Terminé
📄 page du résumé + affichage du résultat:	4 décembre 2025 16:00 → 17	F FARAH Mohamed Walid	Terminé

# Méthodes d'équipe

- **Complexity matching (Sprint 3)** : tâches NLP difficiles aux personnes les plus à l'aise.
- **Pair programming** : utile sur similarité / résumé (Driver–Navigator).
- But : réduire les erreurs et livrer plus vite dans un sprint court.

# Rétrospective : déroulement du projet

- Début : on a sous-estimé certaines tâches (ex. TF-IDF Wikipédia).
- Milieu : meilleure organisation (répartition + binômes)  $\Rightarrow$  meilleure stabilité.
- Fin : on a consolidé snippets + ponctuation.

# Difficultés rencontrées

- Estimation : TF-IDF Wikipédia plus long que prévu.
- Intégration front/back : calcul OK mais affichage pas prêt (page/menu).
- Tests : peu de tests bout-en-bout (temps de séance).
- Organisation : absences et répartition parfois déséquilibrée.

# Leçons apprises

- Garder 10 min fin de sprint : **merge + test complet + mini démo.**
- Une DoD claire aide : *affiché et testable > juste calculé.*
- Complexity matching + pair programming : bon choix pour livrer vite.

# Perspectives

- Stabiliser l'app : version propre et intégrée (pages + modules).
- Ajouter des tests simples bout-en-bout : texte → résultat affiché.
- Améliorer la ponctuation