HELIOS

Projet Inter-Promo SID 2021

SOLUTIONS BY

Berger Levrault

Agenda

1. Quelques généralités
2. Présentation du projet
3. Groupes
4. M2 non-alternants

Généralités

- Les 3 promotions impliquées (≈ 110 étudiants)
- Une thématique commune proposée par Berger-Levrault
- 8 groupes de 10/15 étudiants
  - Un enseignant référent par groupe
  - Un chef de groupe + adjoint
  - Un responsable qualité par groupe (+++++)
- Inscription à Moodle (EIMAB4B1) avec clé : interpromo2019
- Questions à adresser à : pitarch@irit.fr, sauvagnat@irit.fr, hubert@irit.fr ou cecile.chouquet@math.univ-toulouse.fr

Projet interpromo et COVID-19

- À ce stade, difficile d'y voir clair
- Scénario le plus probable : distanciel
- Importance accrue de :
  - La partie gestion de projet
  - La compréhension globale par tous les membres du projet de son rôle
  - L'outil de communication choisi (Discord)

Dates clés

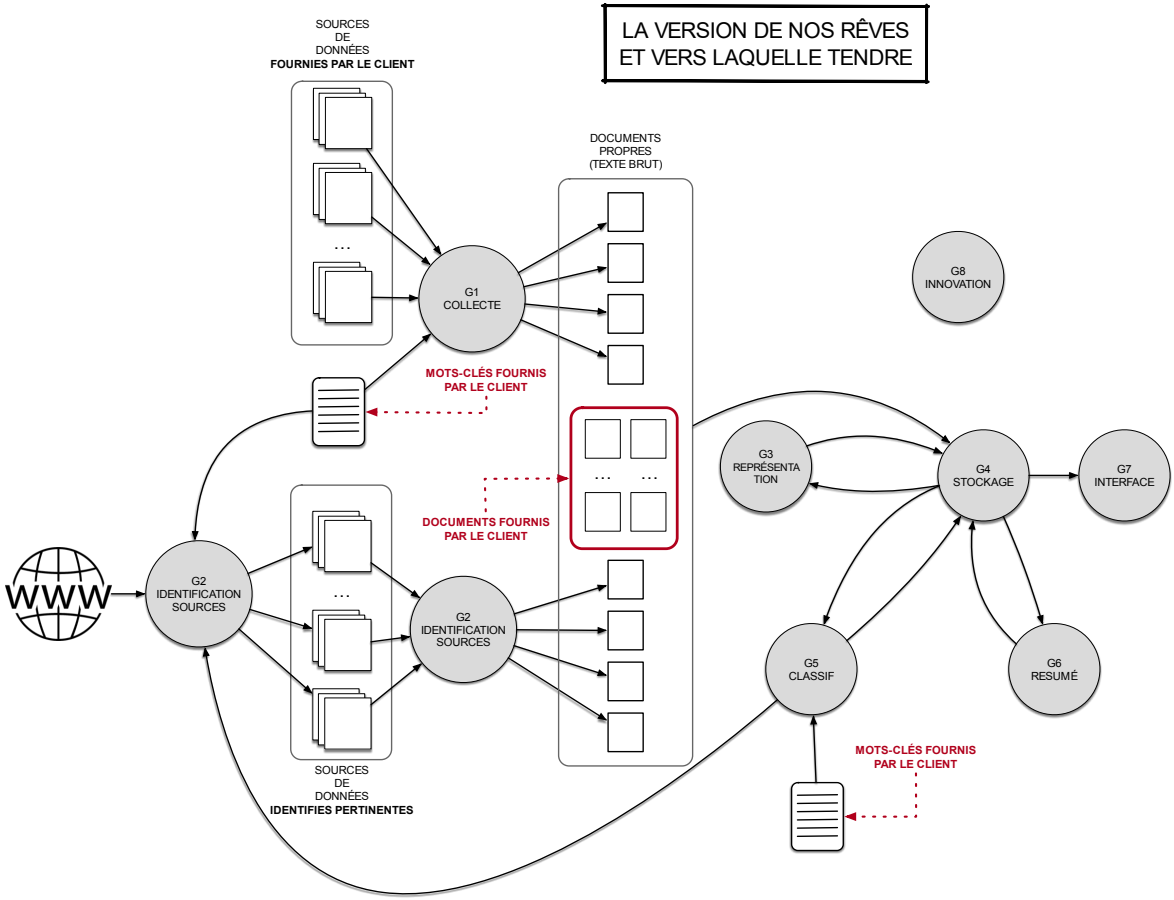
- Phase préparatoire
  - Composition et description des groupes : le 16/11 à 21h sur Moodle
  - Formation gestion de projets en distanciel (Catalyseur) : 25/11 matin (M2), 01/12 matin (M2) et 11/12 après-midi (tous)
  - Accès aux données
    - Données brutes : le 14/12 à 8h
  - Travail préparatoire des M2 non-alternants : du 14/12 au 18/12
  - Travail préparatoire : du 16/11 au 18/12
- Période du projet : 04/01 au 15/01
- Restitution du projet : le 18/01 après-midi

Présentation du projet

Objectif principal

Construire un observatoire des innovations autour des métiers de la gestion.

- ➔ Récolte de données textuelles
- ➔ Apprentissages divers sur ces données
- ➔ Construction/visualisation de résumés et de statistiques sur le corpus


LA VERSION DE NOS RÊVES
ET VERS LAQUELLE TENDRE
SOURCES
DE
DONNÉES
FOURNIES PAR LE CLIENT
DOCUMENTS
PROPRES
(TEXTE BRUT)
G1
COLLECTE
G8
INNOVATION
MOTS-CLÉS FOURNIS
PAR LE CLIENT
G3
REPRÉSENTA
TION
G4
STOCKAGE
G7
INTERFACE
DOCUMENTS FOURNIS
PAR LE CLIENT
WWW
G2
IDENTIFICATION
SOURCES
G2
IDENTIFICATION
SOURCES
G5
CLASSIF
G6
RESUMÉ
SOURCES
DE
DONNÉES
IDENTIFIES PERTINENTES
MOTS-CLÉS FOURNIS
PAR LE CLIENT

Travail préparatoire*

- TOUT le travail réalisé pendant la formation gestion de projet (brown papers, estimation de la durée des tâches et prévision des risques)
- Tous les membres d'un groupe doivent avoir compris le travail à accomplir et les tâches à réaliser
- Librairies Python pour la data science (NumPy, Pandas, Scikit-learn, …)
- Chartes de codage
- Tout travail supplémentaire demandé par le chef de groupe ou l'enseignant référent

* tous les étudiants sont concernés par ce travail préparatoire

Les groupes

Groupe 1
Collecte de données

Adrien Mazoyer

Objectifs

Au sein de sources de données fournies par le client, rechercher les articles relatifs à certains mots-clés fournis par le client puis scrapper ces documents.

Difficultés :

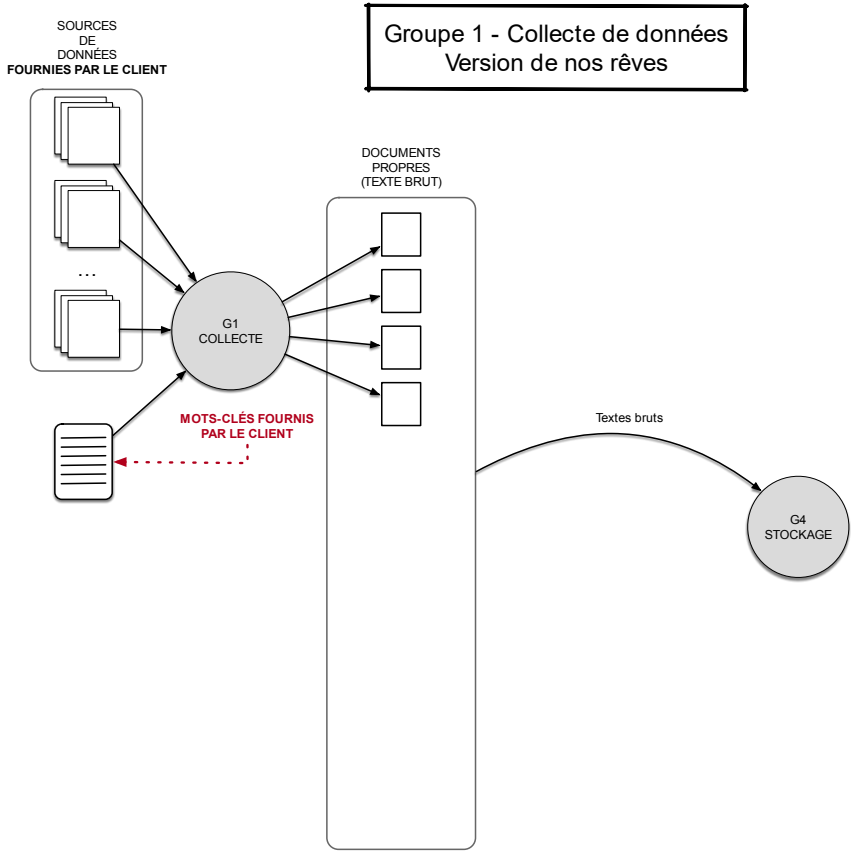
De nombreux mots-clés => priorisation par source

- Identifier les doublons syntaxiques

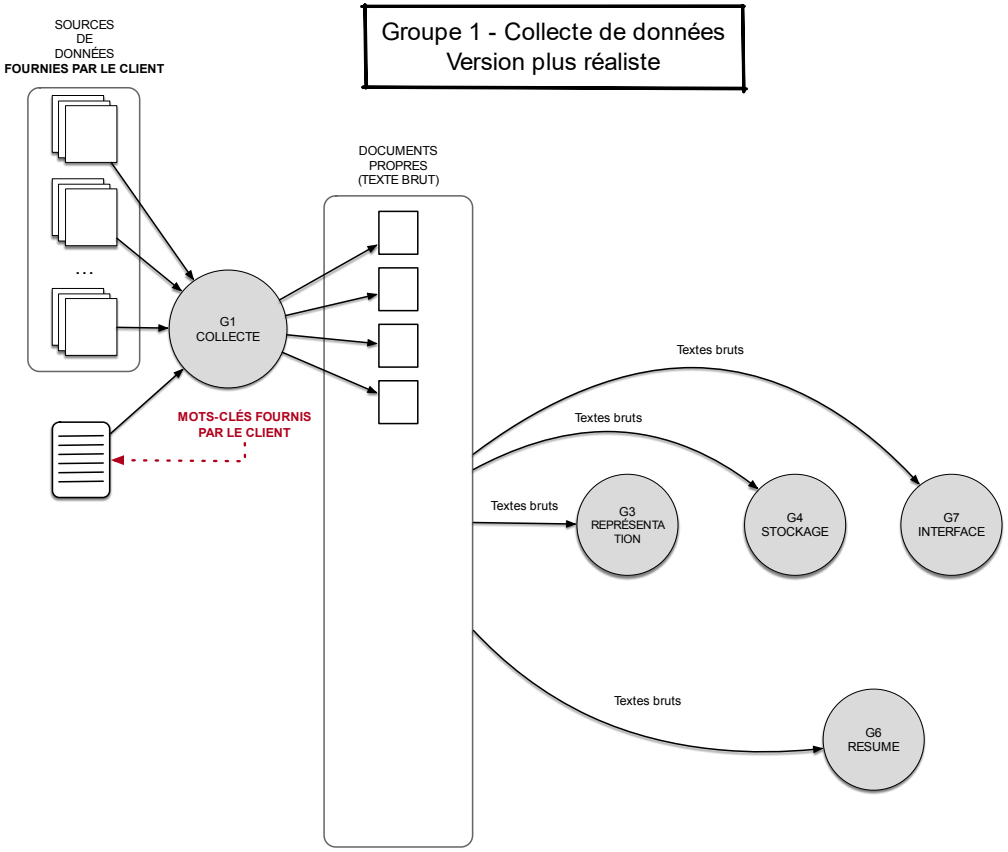
Délivrables

- Données documentées
- Scripts de recherche / priorisation / collecte commentés
- Un rapport descriptif par source (volume, organisation des répertoires, lancement des scripts, mise en place de crown, processus de reprise sur panne, …)

G1 – V2 – INPUT / OUTPUT

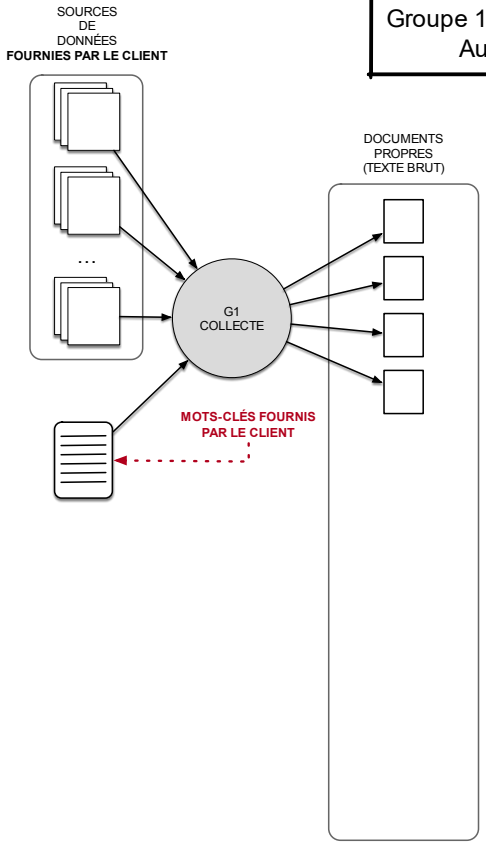


G1 – V1 – INPUT / OUTPUT



G1 – V0 – INPUT / OUTPUT

Groupe 1 - Collecte de données
Au début du projet



Groupe 2
Identification et scrap de nouvelles sources

Yoann Pitarch

Objectifs

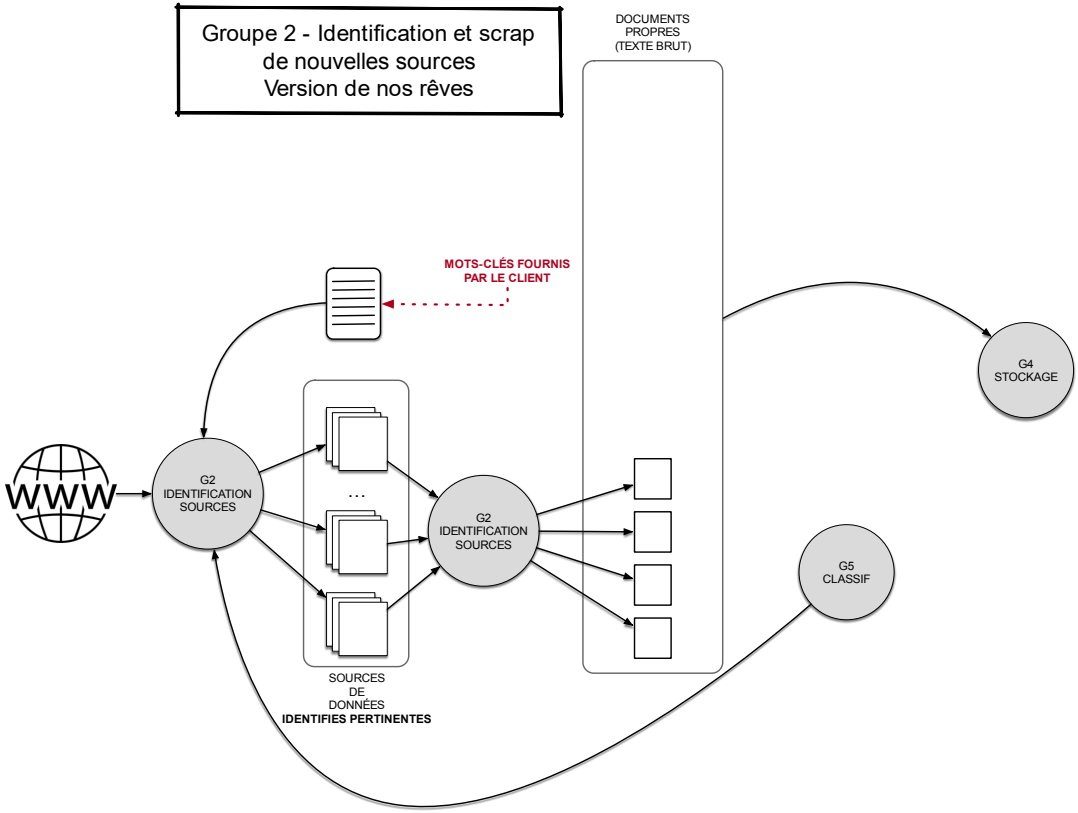
Identifier de nouvelles sources de données contenant des articles pertinents pour le client et scrapper ces articles pertinents.

Difficulté : implémentation d'un scrapper générique

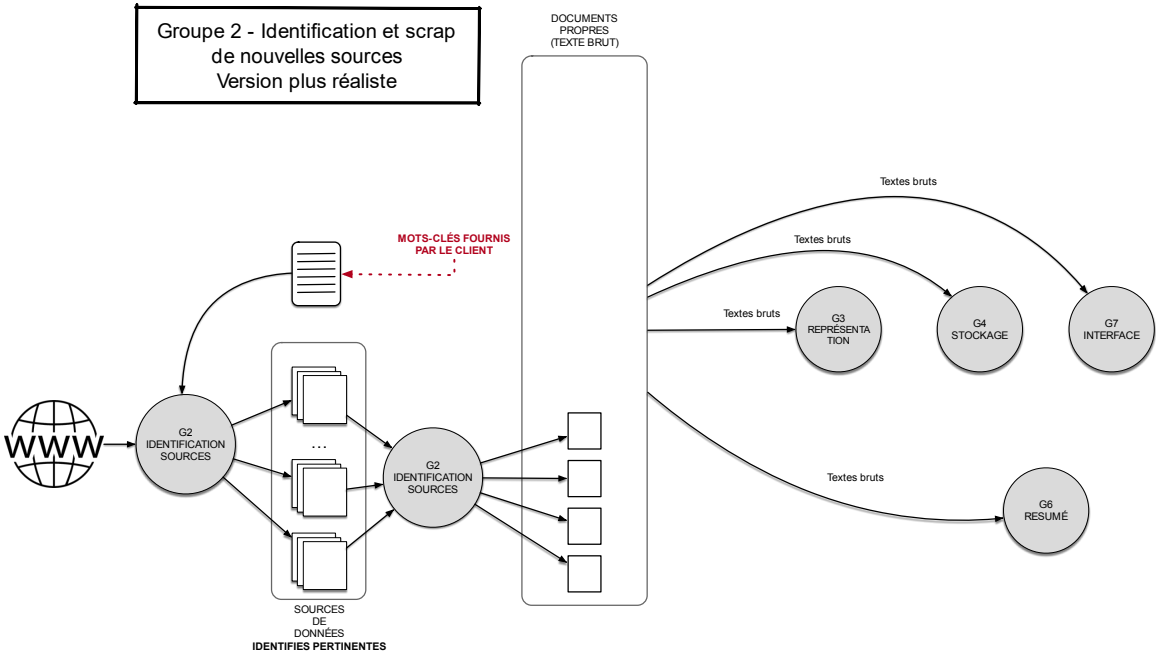
Délivrables

- Données documentées
- Scripts de collecte commentés
- Notebook détaillé sur la création de modèles et la découverte de faisceaux de preuves pour établir la priorisation

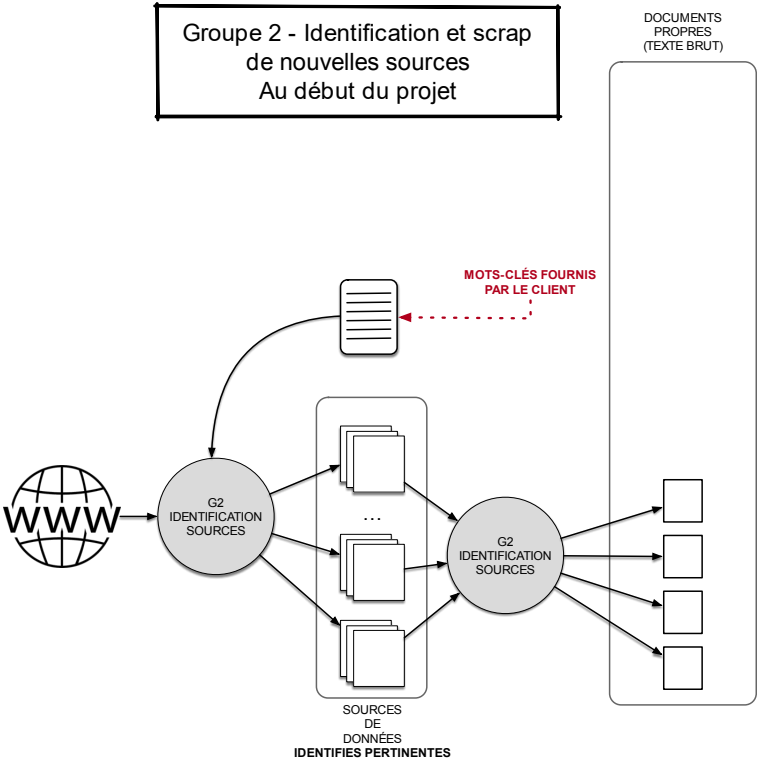
G2 – V2 – INPUT / OUTPUT



G2 – V1 – INPUT / OUTPUT



G2 – V0 – INPUT / OUTPUT



Groupe 3 Représentation de documents

Alexis Dusart

Objectifs

Produire différentes représentations des textes bruts (et des méta-données) pour alimenter les différents groupes d'analyse. Exemple de représentations : sac de mots, embeddings, mots-clés, topics, …

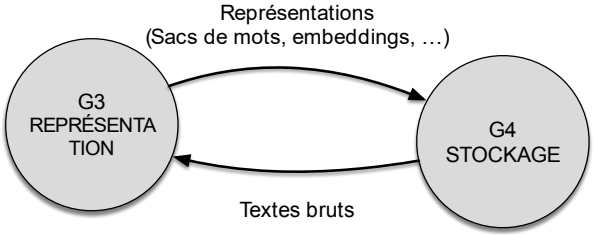
Difficulté : interactions fréquentes et nécessaires avec les groupes d'analyse

Délivrables

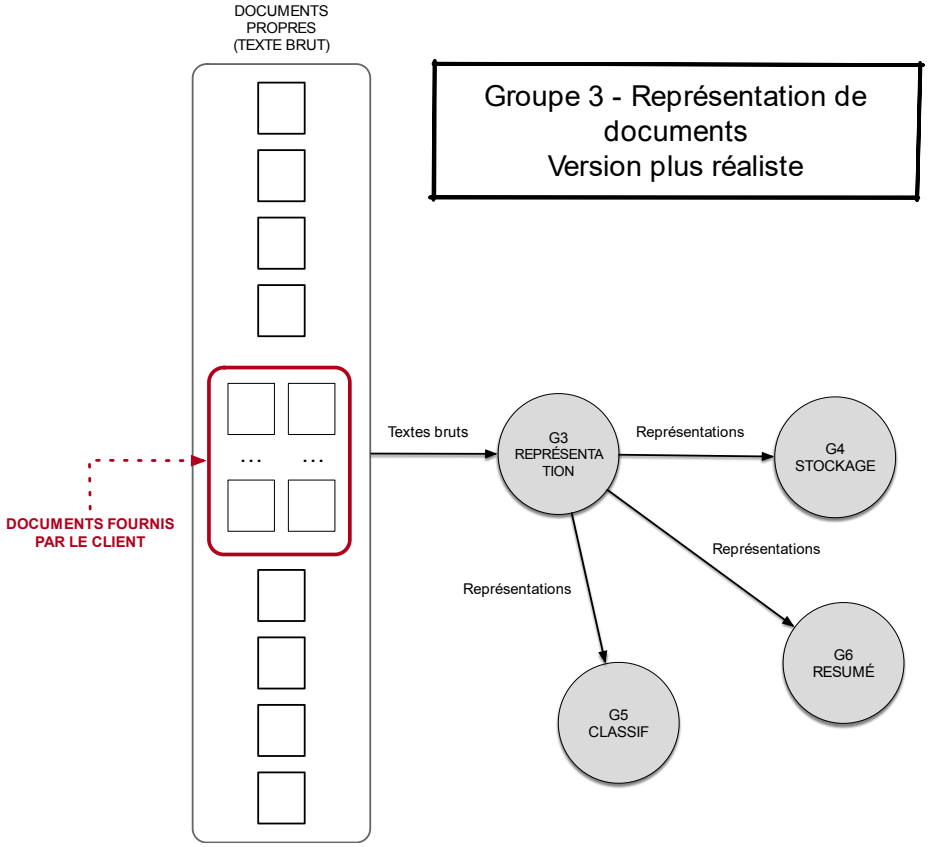
- Les codes de génération des représentations (notebooks)
- Les différentes représentations
- Un document de documentation décrivant les différentes représentations

G3 – V2 – INPUT / OUTPUT

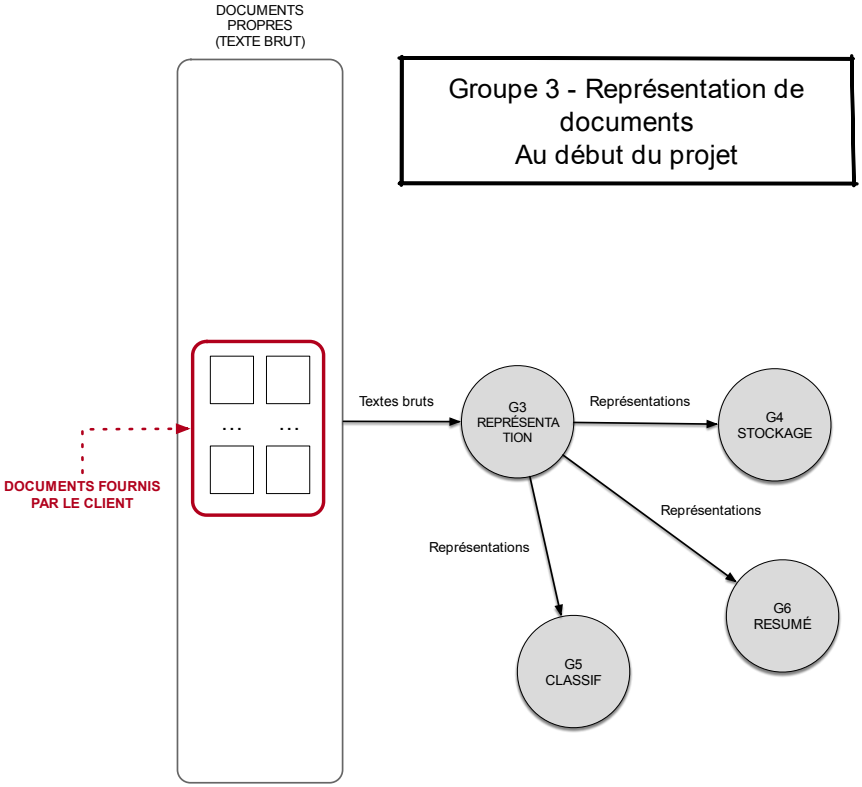
Groupe 3 - Représentation de documents
Version de nos rêves



G3 – V1 – INPUT / OUTPUT



G3 – V0 – INPUT / OUTPUT



Groupe 4
Stockage

Gilles Hubert

Objectifs

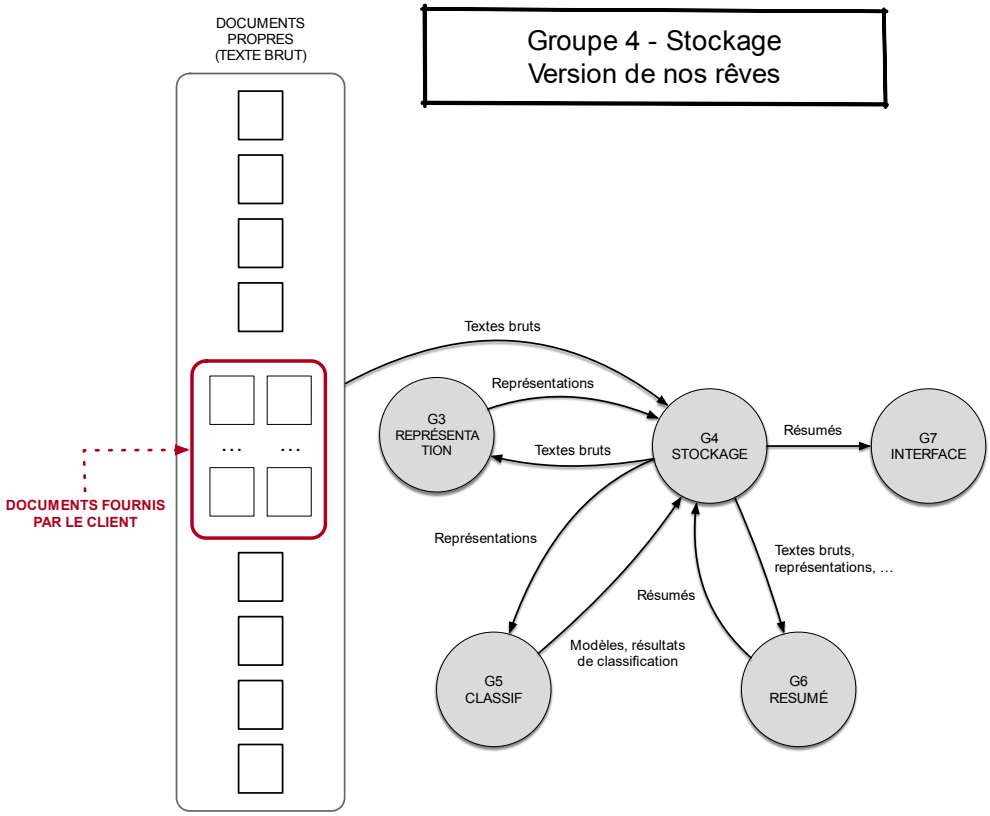
Proposer des schémas de stockage de toutes les données manipulées et générées. Un ou plusieurs SGBD (très probablement NoSQL) seront utilisés en fonction des besoins.

Difficulté : proposer une méthode d'accès aux données en lecture/écriture (API)

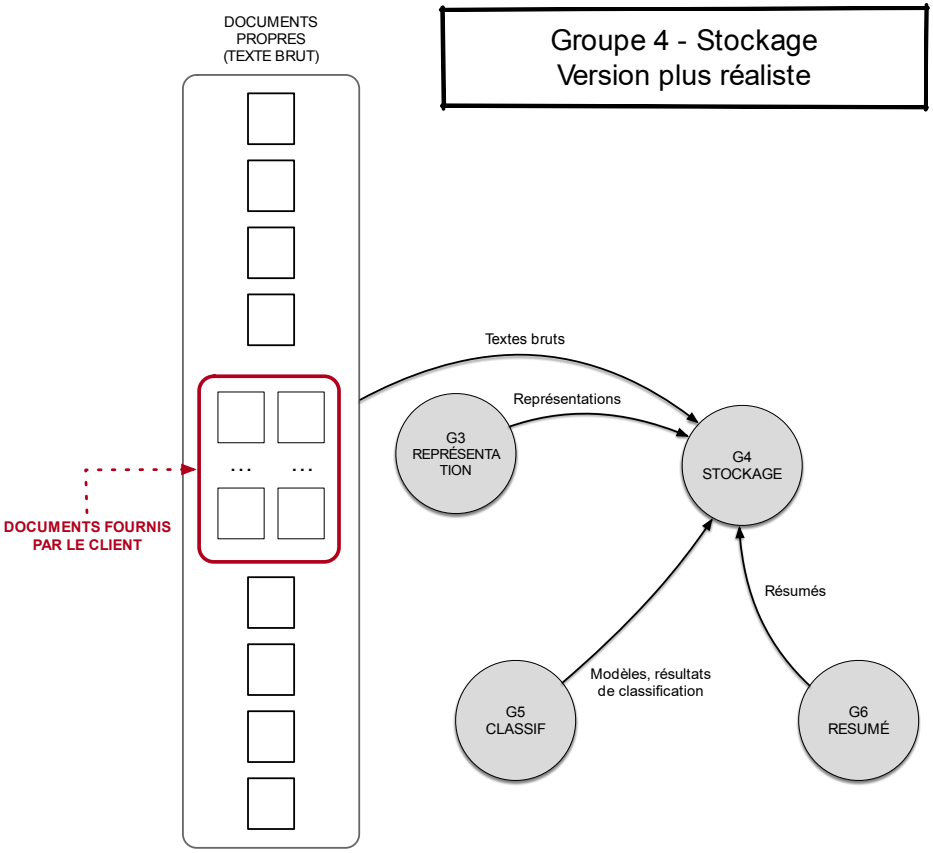
Délivrables

- Benchmark des solutions possibles en fonction des besoins
- Documentations sur le ou les modèles de données choisis
- Tous scripts d'insertion / test / interrogation et toute documentation utile associée
- Dump des données à la fin du projet (documentée)

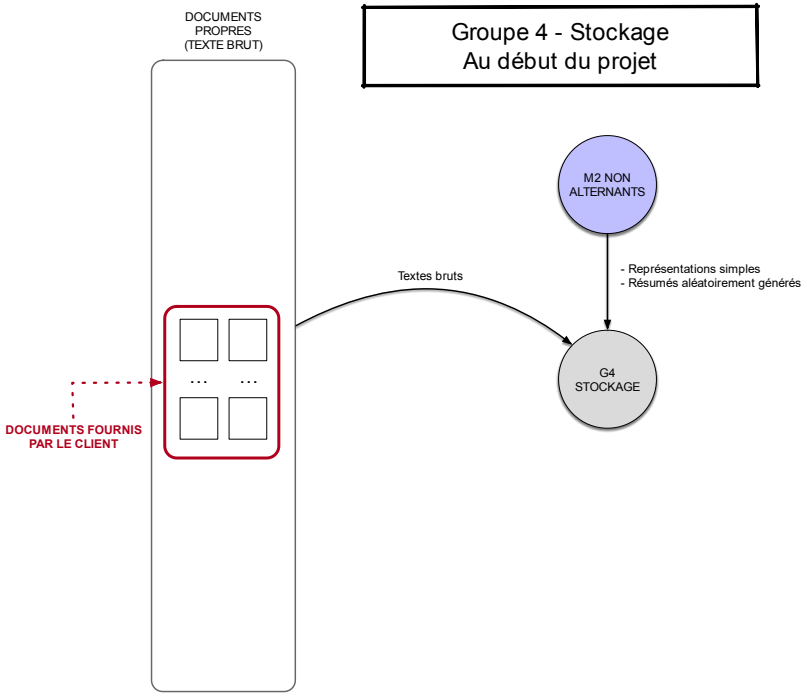
G4 – V2 – INPUT / OUTPUT



G4 – V1 – INPUT / OUTPUT



G4 – V0 – INPUT / OUTPUT



Groupe 5
Classification

[redacted]

Mathieu Serrurier

Objectifs

Différents apprentissages devront être réalisés :

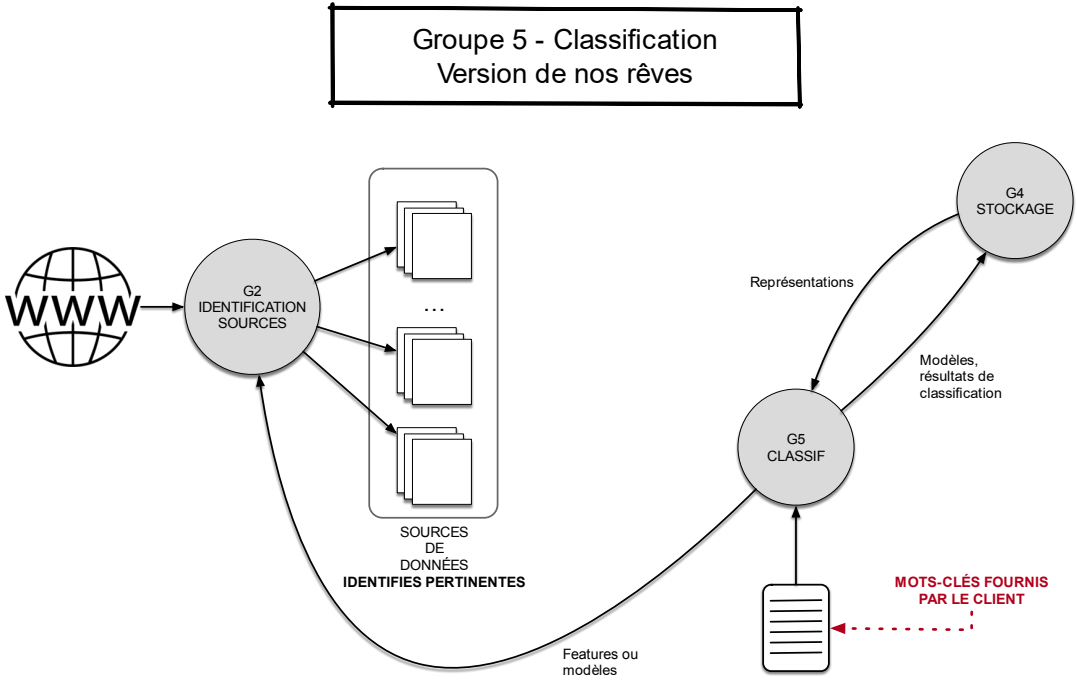
- INNOVATION OU NON : détermine si un document décrit une innovation (et est donc intéressant à analyser plus finement).
- THÉMATIQUE : regroupe les documents autour de leur thématique pour création de *stories*
- NOUVEAUTÉ OU NON : détermine si un document apporte une nouvelle information par rapport aux documents déjà extraits (détection de doublons sémantiques)

Difficulté : évaluation !!

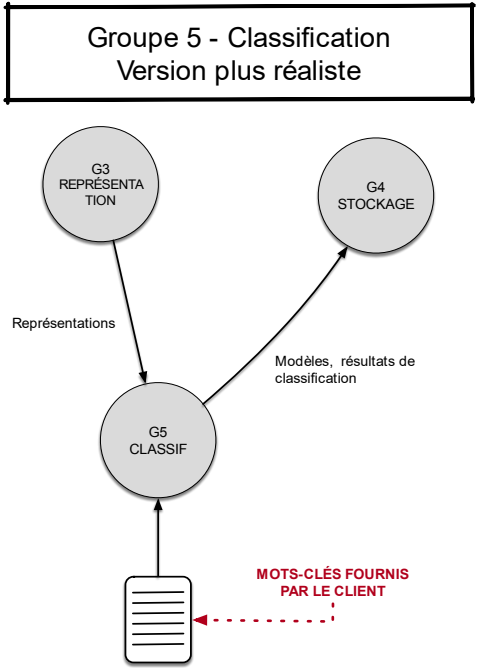
Délivrables

- Notebooks associés à toutes les tâches d'apprentissage effectuées ainsi que leurs évaluations

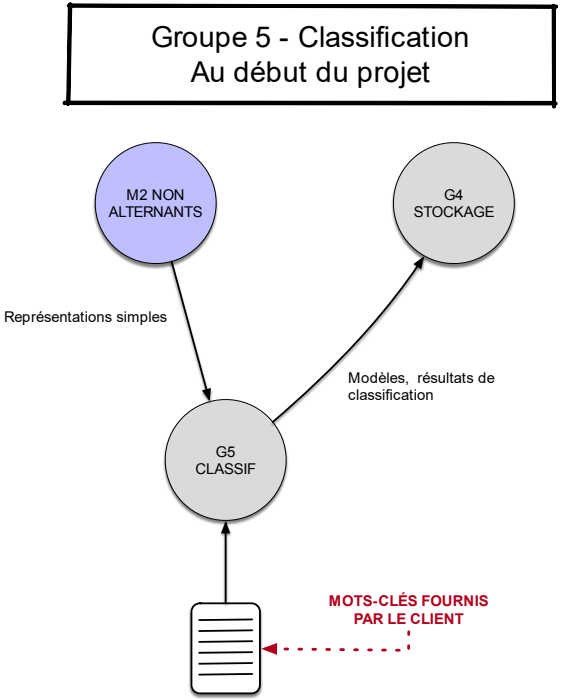
G5 – V2 – INPUT / OUTPUT



G5 – V1 – INPUT / OUTPUT



G5 – V0 – INPUT / OUTPUT



Groupe 6
Résumés

Karen Pinel-Sauvagnat

Objectifs

Deux niveaux d'agrégation sont envisagés :

- Résumé de document (semaine 1) : application de méthodes de résumé automatique sur un seul document. (semaine 1)
- Construction d'un agrégat : construction d'un résumé multi-documents pour chaque story. Le résumé devra représenter tous les aspects de la story et minimiser la redondance. (semaine 2)

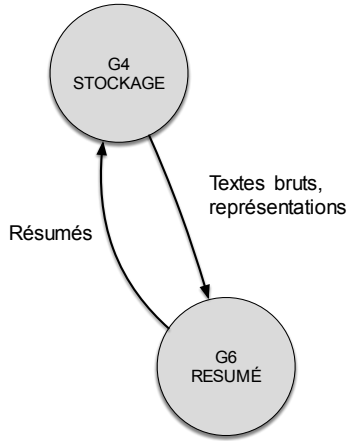
Difficulté : évaluation !!

Délivrables

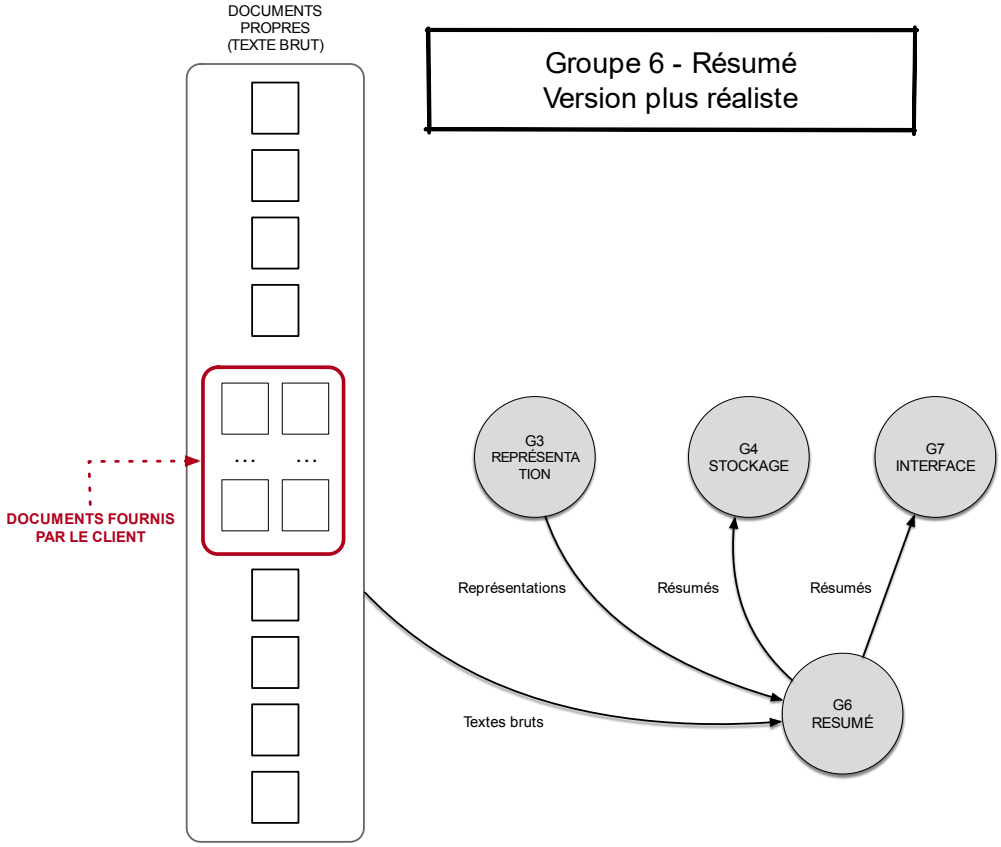
- Notebooks associés à toutes les tâches de résumé effectuées ainsi que leurs évaluations

G6 – V2 – INPUT / OUTPUT

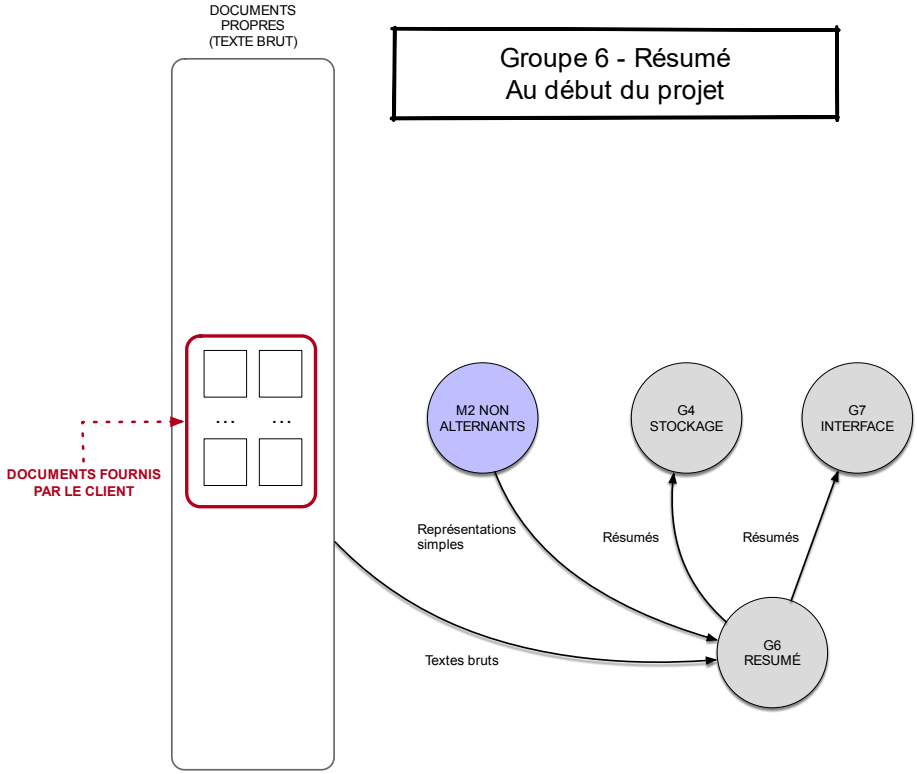
Groupe 6 - Résumé
Version de nos rêves



G6 – V1 – INPUT / OUTPUT



G6 – V0 – INPUT / OUTPUT



Groupe 7
Visualisation

Cécile Chouquet

Objectifs

Trois tâches abordées en parallèle **(ou pas)** :

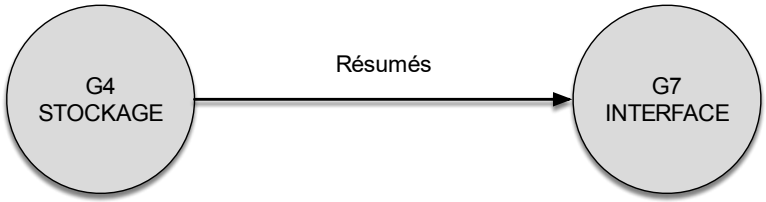
- Visualisation des résumés : cette première tâche est considérée comme la plus importante. La visualisation des résumés est envisagée sous forme d'une timeline, de stories, etc…
- Affichage de statistiques sur le corpus de documents : cette tâche a pour objectif de situer la masse d'informations traitées par l'outil en termes de nombre de documents, de sources, …
- Proposition d'une interface de visualisation interactive des résumés et statistiques

Délivrables

- Logiciel de Dashboard avec code commenté

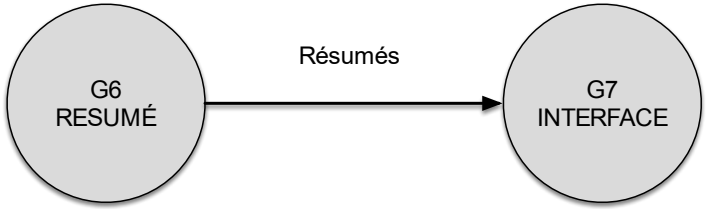
G7 – V2 – INPUT / OUTPUT

Groupe 7 - Interface
Version de nos rêves



G7 – V1 – INPUT / OUTPUT

Groupe 7 - Interface
Version plus réaliste



G7 – V0 – INPUT / OUTPUT

Groupe 7 - Interface
Au début du projet

Groupe 8
Innovation

[redacted]

José Moreno

Objectifs

L'objectif de ce groupe est d'imaginer des axes d'amélioration au travail réalisé par les autres groupes (ajout de fonctionnalités, automatisation, …).

Par exemple :

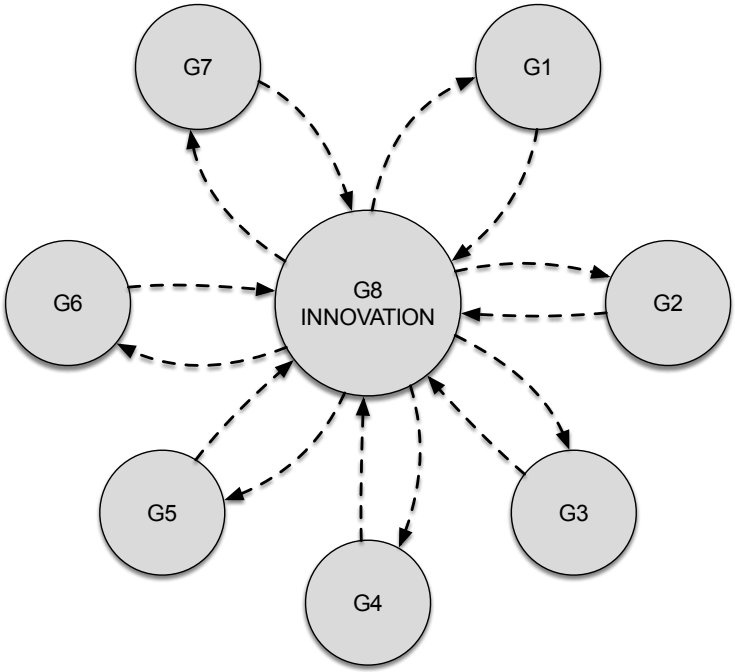
- Dater une story ? (comment définir la période associée à une story)
- Gestion des données multimédia ?
- ...

Délivrables

Le rendu pourra être un POC ou alors une revue de la littérature autour d'un sujet.

G8 – THE WORLD IS YOURS

Groupe 8 - Innovation



M2 non-alternants

Travail à réaliser du 14/12 au 18/12

- ➔ Génération de représentations très peu sophistiquées
- ➔ Génération de résumés / stories aléatoires
- ➔ *Peut-être d'autres choses…*

Questions ?

Dates clés

- Phase préparatoire
  - Composition et description des groupes : le 16/11 à 20h sur Moodle
  - Formation gestion de projets (Catalyseur) : 25/11 matin, 01/12 matin et 11/12 après-midi
  - Accès aux données
    - Données brutes : le 14/12 à 8h
  - Travail préparatoire des M2 non-alternants : du 14/12 au 18/12
  - Travail préparatoire : du 16/11 au 18/12
- Période du projet : 04/01 au 15/01
- Restitution du projet : le 18/01 après-midi