WATCHNEWS

WHO WATCH THE NEWS PAPER?

C. Chouquet, K. Pinel-Sauvagnat, L. Tamine-Lechani, Y. Pitarch, M. Serrurier

Les groupes seront déposés sur le moodle <u>après</u> la présentation

Description du projet

- Site web d'analyse des médias numériques :
  - *Temps réel*
  - *Analyse des tendances thématique*
  - *Analyse sémantique des articles*
  - *Comparaison du traitement d'un thème en fonction des sources*
- But : étudier les biais de traitements des différentes sources (au minimum 15)
- Principes :
  - *Au moins une quinzaine de sources*
  - *Index par mot (le site ne contient pas les articles)*

USE CASE 1: Page de garde

- Liste des sources
- Évolution du nombre d'articles :
  - *Par thème (politique intérieure, politique internationale, sport, santé, high-tech, etc. …)*
  - *Par période (jour, semaine, mois)*
- Les tendances par thème et période
  - *Mot de la journée/semaine/mois*
  - *Verbe de la journée/semaine/mois*
  - *Adjectif de journée/semaine/mois*

USE CASE 1: Exemple

- Sources :
  - *Le Monde*
  - *Le Figaro*
  - *Le Gorafi*
  - *Sciences et avenir ..*
- Statiques
  - *Total articles : 16 000*
  - *Articles jour/semaine/mois:*
    - 40 ↑ ↑ /200↔/1200↓
- Politique internationale
  - *Nom : Catalogne* ↓
  - *Verbe : voter* ↔
  - *Adjectif : indépendance* ↔
- Culture
  - *Nom : Weinstein* ↑ ↑ ↑
  - *Verbe : Agresser* ↑
  - *Adjectif :* sexuel ↑

USE CASE 2 : Recherche par mot-clef

- Pour un mot clef et une source
- Statistique de traitement
  - *Importance du traitement par rapport aux autres sources*
  - *Répartition par thème*
  - *Évolution du traitement dans le temps*
  - *Mot/adjectif/verbe associé*
- Analyse sémantique du traitement
  - *positif/négatif*
  - *Humoristique/sérieux*
  - *Sujet grave/sujet léger*
  - *…*

USE CASE 2 : exemple Trump/le monde

- Statistique de traitement
  - *1 articles sur 100 (rang : 2)*
    - Politique internationale : 70%
    - Politique nationale : 2%
    - Culture : 0.1%
  - *Jour ↔/semaine ↔/ mois ↑*
  - *Guerre/incompétent/menacer*
- Analyse sémantique du traitement
  - *Négatif 90% ↑*
  - *Humoristique 45 % ↔*
  - *Sujet grave 75% ↑ ↑*

USE CASE : précisions

- L’exemple est une façon possible de présenter les choses (les idées sont les bienvenues)
- L’information doit être présentée de façon numérique et graphique quand c’est possible
- On doit pouvoir afficher ces informations pour une source donnée (use case 1)
- On doit pouvoir comparer ces informations pour deux sources. Quelques idées :
  - *Donner le rang mot de la semaine d’une source pour une autre source (use case 1)*
  - *Indiquer quand les différences sont statistiquement significatives*

Livrables

- Le site web
- La base de données
- Le bot de récupération des données
  - *Récupération*
  - *Filtrage*
  - *Analyse*
- Le module d’analyses statistiques
- Les rapports

Moyens

- Sont fournis : environ 16000 articles issues des sources suivantes :
  - *Le monde*
  - *Le figaro*
  - *Libération*
  - *Le nouvel observateur*
  - *Télérama*
  - *Futurascience*
- Un serveur de test et un serveur d’exploitation (ouvert sur l’extérieur)
- Certaines technologies seront imposées (ex : php/mysql)

Organisation

- Equipe projet inter promo (C. Chouquet, K. Pinel-Sauvagnat, L. Lechani, Y. Pitarch, M. Serrurier)
- 10 Groupes avec un enseignant référent
- Travail pour chaque groupe
  - *Travail préliminaire*
  - *Livrable*
  - *Rapport/ soutenance*
- Notation
  - *Une note individuelle de travail*
  - *Note travail préliminaire*
  - *Note livrable*
  - *Note Rapport*
  - *Note soutenance*

Calendrier

- Rapidement : réunion avec les chefs de groupe
- 15 décembre : rendu travail préliminaire
- 21-23 décembre : M2 serveur 0
- 8-18 janvier projet inter promo
- 19 janvier : soutenance

Groupe 1 : supervision

- Enseignant référent: Mathieu Serrurier
- Objectif : Supervision du choix, questions juridiques : Droits sur la récupération / stockage de contenu.
- Travail préliminaire :
  - *Outils de communications*
  - *bilan des outils et technos, planning, schéma initial du projet*
  - *Format de communication*
  - *Choix des sources (avec abonnement possible)*
- Délivrable : présentation finale du projet, mise à jour du schéma de l'architecture du projet complet, Déploiement sur le serveur de production

Groupe 2 : BD-index

- Enseignant référent: Riad Mokadem
- Responsable : [redacted]
- Objectif : Créer et générer le remplissage de la base de données indexant les mots des articles. Gérer l'évolution de la base (lié aux différentes analyses). Gestion des log de recherche
- Travail préliminaire :
  - Schéma v0 de la base
  - *Planning*
  - *Gestion de l'évolution de la base*
- Délivrable : présentation finale du projet, mise à jour du schéma de l'architecture du projet complet, Déploiement sur le serveur de production
- Technologie : Php/mysql
- Note : les groupes 2 et 3 devront travailler ensemble pour la mise en place initiale de la base

Groupe 3 : BD-Serveur

- Enseignant référent: Shaoyi Yin
- Responsable : [redacted]
- Objectif : Gérer les requêtes d'interrogation entre le site et la base. Gérer aussi l'interface avec l'analyse statistique
- Travail préliminaire :
  - Schéma v0 de la base
  - *Planning*
  - *Format de réponse au site*
  - *Format de requête à l'analyseur statistique*
- Délivrable : serveur PHP d'interrogation
- Technologie : PHP/mysql
- Note : les groupes 2 et 3 devront travailler ensemble pour la mise en place initiale de la base

Groupe 4 : Robot

- Enseignant référent: Yoann Pitarch
- Responsable : [redacted]
- Objectif : L'objectif de ce groupe est d'implémenter un robot qui récupère le contenu des articles de journaux en ligne
- Travail préliminaire :
  - *bilan des outils (choix du langage et des librairies utilisés pour accéder au contenu d'un site internet et la récupération de son contenu)*
  - *appropriation du langage HTML.*
  - *Format d'échange avec le groupe filtrage*
  - *Planning*
- Délivrable : Rapport, Cron robot (au moins 15 sources)

Groupe 5 : Filtrage de texte

- Enseignant référent: Gille Hubert
- Responsable : [redacted]
- Objectif : Filtrage, lemmatisation, mettre les textes en forme, trouver les mots composés, le tout avec un temps de traitement acceptable :-)
- Travail préliminaire :
  - *Bilan des techniques*
  - *bilan des outils*
  - *Format d'échange avec le groupe analyse sémantique*
  - *Planning*
- Délivrable : Rapport, Cron filtrage

Groupe 6 : Analyse sémantique

- Enseignant référent: Lynda Tamine-Lechani
- Responsable : [redacted]
- Objectif : d’enrichir les index de base par une couche sémantique
- Travail préliminaire :
  - *Bilan des techniques*
  - *bilan des outils*
  - *Format d’échange avec le groupe prédiction*
  - *Planning*
- Délivrable : Rapport, Cron analyse sémantique

Groupe 7 : Prédiction / Classification

- Enseignant référent: Sébastien Gerchinovitz
- Responsable : [redacted]
- Objectif : Prédire la catégorie de l'article (santé, culture, politique, sport…). analyse non supervisée à partir des indicateurs sémantiques
- Travail préliminaire :
  - *Bilan des techniques*
  - *bilan des outils*
  - *Format d'échange avec le groupe la BD*
  - *Planning*
- Délivrable : Rapport, Cron filtrage

Groupe 8 : Analyse de tendance/ statistiques

- Enseignant référent: Cécile Chouquet
- Responsable : [redacted]
- Objectif : Analyse de tendance temporelle pour un mot, traitement d'un mot au cours du temps. Comparaison statistique sources
- Travail préliminaire :
  - *Bilan des techniques*
  - *Bilan des outils*
  - *Analyse statique/ dynamique*
  - *Format d'échange avec le groupe la BD*
  - *Planning*
- Délivrable : Rapport, Serveur analyse statistique

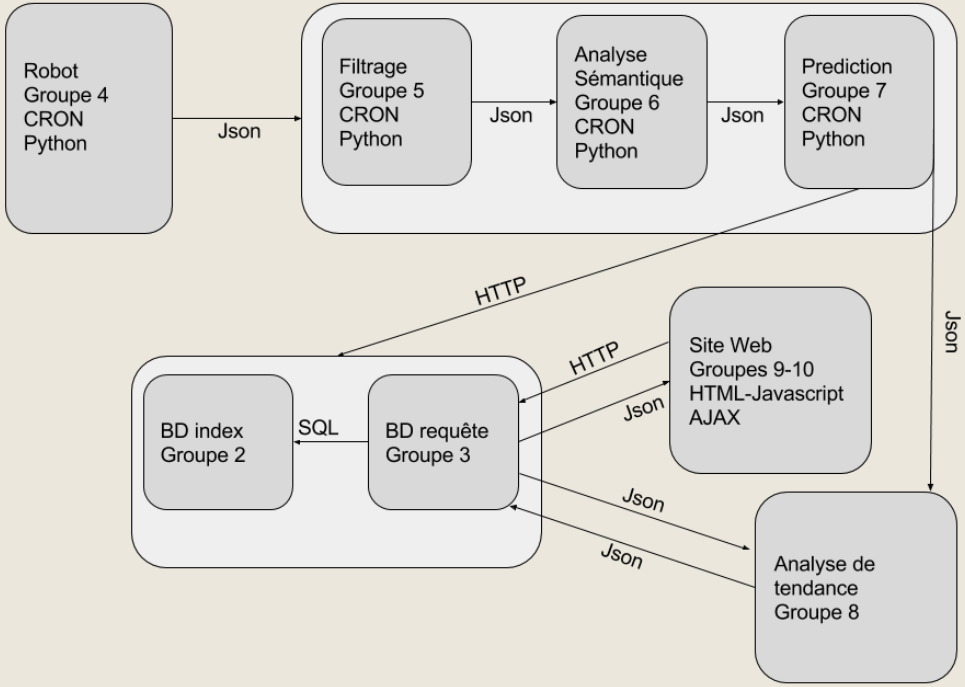
Groupe 9 : Application Web

- Enseignant référent: Karen Pinel-Sauvagnat
- Responsable : [redacted]
- Objectif : Site web responsive avec visuel décidé par groupe 10.
- Travail préliminaire :
  - *Bilan des outils (AJAX,CSS)*
  - *Format d'échange avec le groupe la BD*
  - *Planning*
- Délivrable : Rapport, Site Web

Groupe 10 : Qualité/Communication

- Enseignant référent: Wahiba Bahsoun
- Responsable : [redacted]
- Objectifs :
  - *Licence sur le code ? Étude > proposition > décision*
  - *Visuel du site web/ statistiques*
  - *Charte de codage + template des rapports à rendre*
  - *Mise en place des tests*
  - *Communication*
- Travail préliminaire :
  - *Bilan des outils (javascript ,CSS)*
  - *Charte de codage + template des rapports à rendre*
  - *Planning*
- Délivrable : Rapport, logo, CSS

Serveur 0


Robot
Groupe 4
CRON
Python
Json
Filtrage
Groupe 5
CRON
Python
Json
Analyse
Sémantique
Groupe 6
CRON
Python
Json
Prediction
Groupe 7
CRON
Python
HTTP
Site Web
Groupes 9-10
HTML-Javascript
AJAX
Json
HTTP
Json
BD index
Groupe 2
SQL
BD requête
Groupe 3
Json
Json
Analyse de
tendance
Groupe 8

Mise au Point

- Il vaut mieux un site limité qui fonctionne qu’un site complexe qui ne marche pas
- L'équipe de direction tranche toute les questions importantes
- Chaque groupe est responsable (avec l’aide du groupe de direction) du déploiement sur le serveur de test.
- Seul le groupe de direction déploie sur le serveur de production (le site final + la communication sur le projet).