Problème de départ

La Feuille Officiel Suisse du Commerce (www.fosc.ch) regroupe quotidiennement un grand nombre de publications officielles (registre du commerce, faillites, poursuites, etc.) de toute la Suisse.

But du projet

- Qui est impliqué dans une entreprise ?

- Dans quelle entreprise est impliquée une personne ?

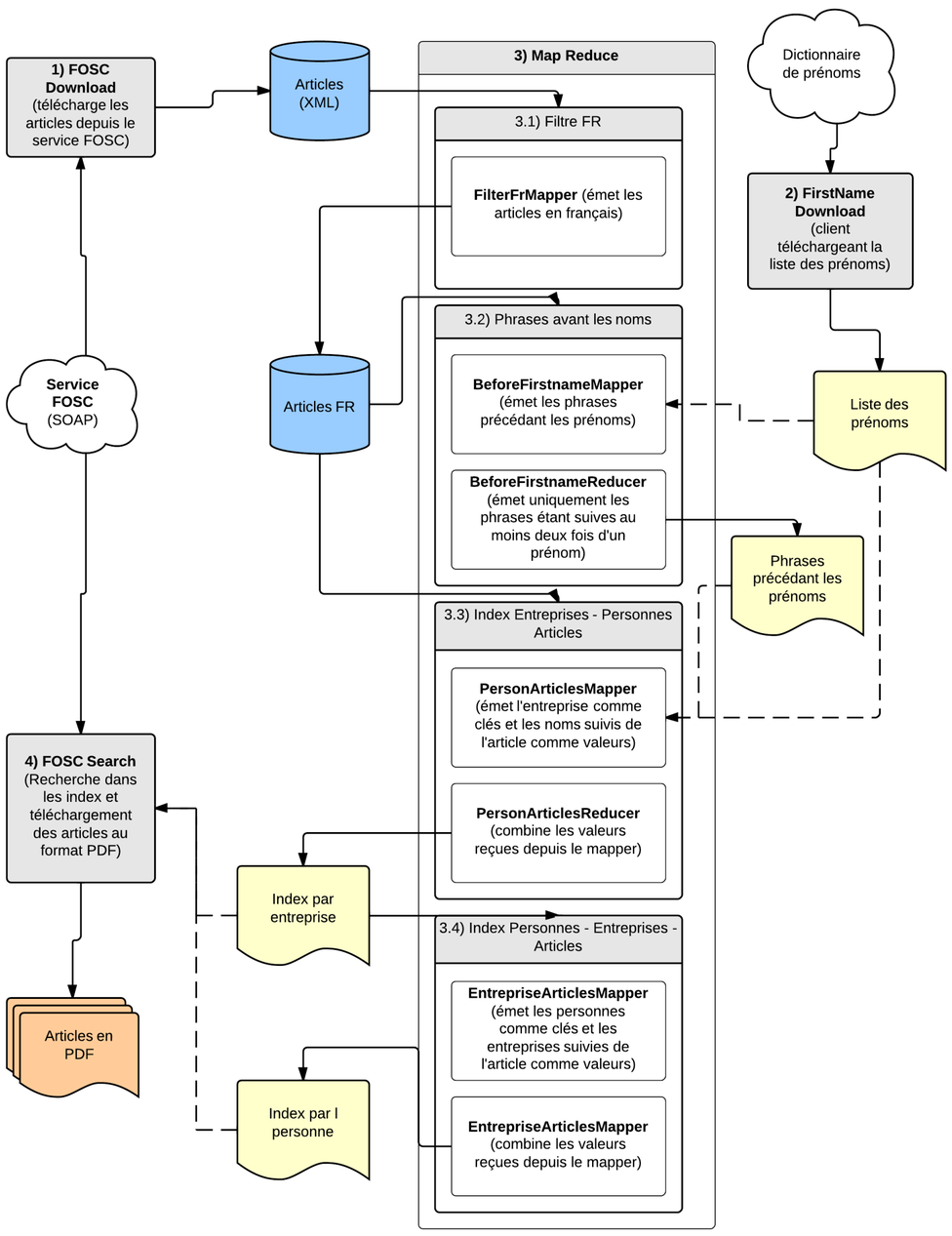
- Qu'elles sont les articles liés à une entreprise ?

- Qu'elles sont les articles liés à une personne ?

Base de données

L'état propose un API (soap) pour récupérer les articles au format XML ou PDF. Nous utilisons le format XML pour indexer les articles et le format PDF pour les afficher. La base de données contient des articles dans les trois langues nationales.

Algorithme (Map / Reduce)



Filtre FR

Seuls les articles en français sont pris en compte. Pour réaliser ceci, un premier « mapper » lit tous les articles et émet uniquement les articles en français.

Phrases avant les noms

Retrouvé le nom de l'entreprise est assez facile. Il est toujours dans le champ <name> de l'article. Par contre, les noms des personnes apparaissent n'importe où dans le texte. Ainsi, la première difficulté consiste à détecter les noms des personnes du reste du texte. Les séquences de mots précédant les noms sont souvent identiques.

Ainsi, dans un premier temps une liste des prénoms (téléchargée depuis « dictionnaire-prenom.bebevallee.com ») est utilisée pour détecter quelles sont les séquences de mots précédant un nom.

Index Entreprises, Personnes, Articles

Lors d'un troisième traitement on retrouve tous les noms à partir des séquences de mots précédant les noms et de la liste des prénoms. On émet un index par entreprises, personnes et articles.

Index Personnes, Entreprises, Articles

Grâce à l'étape précédente on est capable de retourvé les personnes et articles liés à une entreprise. Pour retrouvé l'inverse (entreprises et articles lié à une personne), le résultat de l'étape précédente est lu. Le mapper émet simplement comme clés le nom des personnes et comme valeurs les entreprises et articles.

Quelques problèmes rencontrés et leurs solutions

- Les prénoms et noms ne sont pas toujours présents dans le même ordre dans les articles. Pour comparer les noms nous stockons chaque partie dans un Hashset, ce qui permet d'ignorer la séquence d'apparition

  Le format de sortie des index devait être dans un format à la fois lisible par l'homme et par l'ordinateur. Pour résoudre ce problème nous avons décidé d'utiliser comme format d'output le **TextOutputFormat** et de modifier les fonctions **toString** des classes modèles à fin d'obtenir comme output un fichier avec le format suivant :
  "Enterprise\tPerson1 Names 1:1, 2, 3, 4, 5, 6;Person2 Names 2:1, 2, 3, 4, 5, 6;"
  "Person1 Names1\tEnterprise1:1, 2, 3, 4, 5, 6;Enterprise1:1, 2, 3, 4, 5, 6;"
  Des expressions réguilières ont été utilisées pour améliorer la détection des noms. Ainsi un nom commence obligatoirement par une majuscule, est suivi de plusieurs caractères, d'un éventuel tiret (qui est lui même suivi d'une majuscule et de caractères.)

- Les noms sont parfois précédés de particules (comme « de » ou « von »), certains mots sont toujours suivis de noms (comme Monsieur, Madame ou Maître) et certains mots sont jamais suivi d'un mot (comme Saint ou St). Partant de ce principe, trois listes de mots ont été créées aidant à détectés les noms, des mots ou des villes.

Évaluation du résultat

Le résultat obtenu correspond aux attentes et en particulier le fait d'avoir un format de fichier facilement lisible et d'une taille compacte. Les personnes et les entreprises sont pratiquement toujours détectées correctement. De plus, il y a assez peu de faux positifs.

Améliorations possibles

Grâce au succès obtenu sur cette réalisation, le marge d'optimisation se trouve essentiellement dans :

- L'implémentation d'une version pour supporter les autres langues (Italien, Allemand)
- L'optimisation de l'utilisation des ressources en termes de mémoire utilisée
- L'optimisation du temps de traitement
- La persistance dans un format de fichier plus structuré pour réduire le temps d'extraction
- La création d'un système pour modéliser la reconnaissance des noms dans les différents standards linguistique tels que la différence entre le model Italien et l'allemande

Big Data Analytics - Aurélien Thévoz et Komanda Phanzu