Recherche d'information et traitement automatique des langues

Violaine Prince
Université de Montpellier 2
LIRMM-CNRS

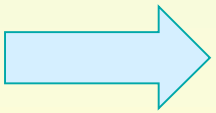
Modèle vectoriel documentaire

- Introduction au modèle de Salton
  - Origine
  - Objectifs
- Document unique
- Base de documents
- Formes complexes de requêtes
- Défauts et limites du modèle

Origine du modèle vectoriel

Modèle de Salton (1968)

- Informatique documentaire
  - Comment classer des documents (indexation)
  - Comment retrouver des documents (recherche d ’information)
  - En fonction des mots qu ’ils contiennent



Fréquence d ’occurrence

Importance des mots

- Hypothèse de Salton :
  - La fréquence d ’apparition d ’un terme en langage naturel est liée à son pouvoir de représentation du contenu de ce texte.
- Indexation automatique :
  - Analyse lexicale des documents et extraction des termes significatifs de leur contenu (1)
  - Pondération des termes pour déterminer leur degré de représentativité (2).

Recherche documentaire selon Salton

- Les termes significatifs créent un « espace », et y sont représentés sous forme de vecteurs.
- **Recherche d'information:**
  - Application d'une « requête » sur un texte.
  - Mesurer la pertinence d'un texte par rapport à une demande.
  - Exemple :
    - Requête :Je cherche un appartement.
    - Le texte T est pertinent pour la requête s'il contient les termes « chercher » et « appartement ».

Recherche documentaire selon Salton

- Pour Salton :
  - Si on est capable d'indexer un texte avec le mot « chercher » et avec le mot « appartement », alors il est pertinent.
  - Recherche d'information et indexation, seraient des processus duaux.
  - Comment procéder au processus d'indexation ?
    - Nature des données
    - Indexation extensive

Premier modèle: document unique

- Extraction des termes représentatifs :
  - Méthode
    - Lemmatiser le texte
    - Récupérer les lemmes -> lexies de dictionnaires
    - Eliminer les lexies dont la catégorie est fonctionnelle : prépositions, conjonction, déterminants, pronoms, etc.
    - Sur le reste: essentiellement les noms communs,les noms propres, les adjectifs, les verbes et les adverbes.
    - Séparer
      - Catégories « nobles » : noms et verbes
      - Catégories auxiliaires : adjectifs et adverbes => retriés (épithète, attributs, adverbes de temps et de lieu).
    - Enlever les mots ordinaires et athématiques (mots d ’usage courant).

Création de la « base vectorielle »

- Les éléments qui restent sont appelés **termes significatifs**.
- Ces éléments sont modélisés comme des vecteurs.
- Il existe plusieurs manières de représenter ces vecteurs.
- **Premier modèle : le modèle booléen**.
  - Les termes significatifs du documents sont projetés chacun sur l'ensemble de la base constituée.
  - Chaque mot m_i est représenté par

    (0,0,0, ….., 1, 0, 0, …, 0) où 1 indique le poids de la composante en ième position (projeté sur lui-même).

Le modèle booléen

- Les vecteurs ainsi créés sont libres.
- Ils sont générateurs par hypothèse.
- Soit B la base vectorielle des mots du document.
- Toute requête R est traitée de la même manière et ne contient que les mots significatifs.
- Ces derniers sont plongés dans la base du document. Le vecteur de la requête est l'union des vecteurs de ses termes significatifs dans B. Il peut être nul.

La pertinence d'une requête

- La **pertinence** d'un texte T de base B, par rapport à la requête R pour est mesurée par le cosinus du vecteur de la requête V(R) avec le vecteur unique union des vecteurs de B, V(B).

Cos(V(R),V(B)) =

(V(R).V(B))/ IIV(R)II IIV(B)

- Où '.' est le produit scalaire de deux vecteurs et II II la norme euclidienne.

Rappels

- Produit scalaire de deux vecteurs
  - $A = (a_1, a_2, \ldots, a_n)$
  - $B = (b_1, b_2, \ldots, b_n)$
  - $A.B = a_1b_1 + a_2b_2 + \ldots a_nb_n$
- Norme d'un vecteur
  - $\|A\| = \sqrt{\sum_i a_i^2}$

Le modèle fréquentiel

- Limites du modèle booléen
  - Tous les mots ont même importance.
  - Or l'hypothèse de Salton va dans le sens: « plus on parle de quelque chose, plus il est discriminant ».
- Introduction du modèle fréquentiel :
  - Les vecteurs de la base ne contiennent plus des valeurs booléennes, mais des valeurs fréquentielles.

Le modèle fréquentiel

- Représentation du vecteur du mot m_i :

$(0,0,\ldots., f(m_i),0,0,\ldots.0)$ où $f(m_i)$ est la fréquence du mot m_i dans le document.

- On conserve la même représentation des requêtes, et la même mesure de la similarité (ou pertinence).
- Résultat :
  - discrimination des valeurs du cosinus.
  - Favorise les requêtes sur les mots très fréquents d'un texte.

Le modèle fréquentiel

- Variantes :
  - Vecteurs normés :
    - On norme les vecteurs de la base.
    - La composante du mot m_i reçoit la valeur
    - $F(m_i)/IIV(B)II$
    - La similarité est alors simple : le cosinus est confondu avec le produit scalaire.
  - Modification de la pondération selon la nature de la catégorie grammaticale.
    - Par exemple, les verbes, les noms communs et les noms propres ont un poids de 1

Le modèle fréquentiel

- Les adjectifs et les adverbes ont un poids de 0,5.
- Chaque composante de la base reçoit la valeur $F(m_i) \times p_i$, normée ou pas.
- Part de l'hypothèse de la notion de gouvernement chez Chomsky: Le nom et le verbe sont gouverneurs dans leur groupe.
- On fait le même traitement pour les termes de la requête.
- Résultat : favorisera les verbes/noms dans les requêtes.
- Contre-exemple : « moyenne pondérée » favorisera « moyenne » alors que le traitement aurait dû être égalitaire. « Médecin de campagne » traitera à égalité « médecin » et « campagne ».
- La notion de « gouvernement » est incomplètement employée.

Premiers constats sur le modèle fréquentiel

- Avantages : mots fréquents mieux discriminés.
- Inconvénients : vecteurs trop grands à cause des mots rares. Beaucoup de bruit.
- Variante : Rétrécissement de la base.
  - On élimine les termes rares (f(ti) < 2).
  - Choix ensuite du modèle : normé, non normé, pondéré par la catégorie grammaticale.
  - Possibilité de variante sur le troisième choix :
    - Rajouter une notion d'**usage** pour affaiblir le poids des mots courants.

Modèle fréquentiel évolué

- Notion de mot courant :
  - Mots joker : truc, bidule, système,objet, chose, entité….
  - Mots « gouverneurs » mais d'usage de liaison :
    - Verbes : faire, être, donner, avoir, etc…
  - On essaie de « border » la pertinence.
  - Techniques plus « langage naturel » :
    - Remplacer des verbes par des noms.
    - Par exemple : Je cherche un appartement, les deux termes restants sont « chercher » et « appartement ».
    - Chercher étant un verbe assez « courant », on le remplace par son substantif ; « recherche ».
    - La requête devient : « recherche » et « appartement ».

Modèle fréquentiel évolué

- Troisième variante : rétrécissement maximal de la base.
  - Au delà des termes rares, on élimine en outre
  - Les mots courants
  - Les termes rares non techniques :
    - Exemple : « superfétatoire », « procrastiner », etc…
  - Conservation des termes « techniques ».
- Tout cela suppose:
  - D'avoir un thésaurus des mots courants
  - De décider quels sont les termes rares non techniques
  - Et quels sont les termes techniques (notion de domaine, de catégorie).
- On voit parfois des modèles fréquentiels associés à des modèles de bases de connaissances (arborescences, réseaux, etc…)

Modèle fréquentiel évolué

- Problèmes :
  - Plus V(B) est petit, plus V(R), vecteur de la requête, risque d'être nul.
  - Introduction de « silence » dans le modèle.
- Avantages :
  - Meilleure précision.
- Conclusion :
  - Les modèles mis en place par Salton et son équipe sur un document unique, ont ensuite été étendus à une **collection de documents**.
  - Toute requête R se fait sur cette collection.
  - L'idée est de récupérer, dans cet ensemble, le ou les documents les plus pertinents par rapport à R.

Extension à une base de documents

- On envisage un ensemble de documents :

  $\{D_1, D_2, \ldots, D_n\}$

- Pour créer la base documentaire :
  - Lemmatisation de tous les documents
  - Suppression des catégories « non signifiantes ».
  - Constitution des vecteurs de termes significatifs, issus de l'union des documents.
  - Création du vecteur de base, V(B), union des vecteurs de la base.

Extension à une base de documents

- Introduction de la notion de **vecteur de document** :
  - $V(D_k)$, vecteur du kième document de la collection, est projeté sur B.
  - Il est composé de l'union des vecteurs des termes significatifs obtenus dans D_k.
  - La valeur de la composante dépend du choix du modèle de représentation.
  - Pour n documents, il y aura donc n vecteurs de documents.
- Le choix du modèle de représentation :
  - Booléen
  - Fréquentiel simple
  - Tf/idf

Base de documents: modèle booléen

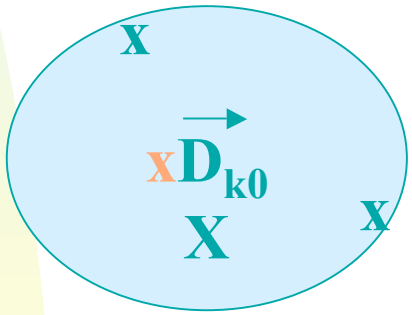
- Ce modèle s'applique non pas à la composante des vecteurs de la base, mais à la composante du vecteur de documents.
- $V(D_k)$ est de la forme
- $(a_1^k, a_2^k, \ldots., a_i^k, a_j^k, \ldots. a_{dimB}^k)$
- Où a_i^k vaut 0 si le terme significatif t_i (i dans {1, dimB}) n'existe pas dans D_k
- 1 sinon.

Base de documents: modèle booléen

- Le vecteur $V(D_k)$ représente les mots de la base utilisés par le document D_k.
- Le modèle permet de comparer les vecteurs de documents entre eux.
- Cette comparaison de documents se fait par la mesure de similarité (ou de dissimilarité) par le cosinus.
- On peut classer les documents d'une collection.

Base de documents: modèle booléen

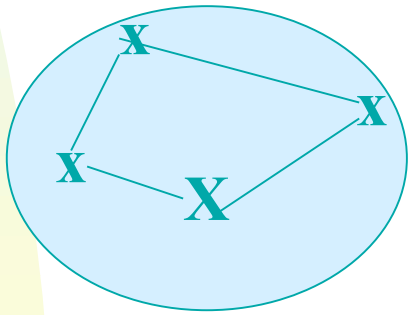
- Méthode de classement :
  - Avec des seuils :
    - Quels sont les documents proche d'un même document avec un cosinus supérieur à un seuil σ ?



Boule de centre $\vec{D}_{ko}$ et de Rayon 1-σ

Base de documents: modèle booléen

- Quels sont les documents qui, entre eux, comparés deux à deux, ont des cosinus de valeur supérieure à σ ?



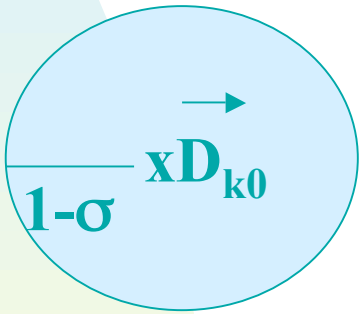
Clique dont les arcs sont de De valeur $1-\sigma$ max.

Comparer des documents et des requêtes

- Classement par valeur relative :
  - Pour toute paire de documents D_i et D_j, on calcule cos ($V(D_i)$, $V(D_j)$).
  - On dira que les plus proches sont les paires qui ont les plus grandes similarités.
- Traitement des requêtes :
  - Toute requête R est lemmatisée et « nettoyée ». Ses termes significatifs sont projetés sur la base B.
  - On produit alors un vecteur V(R) booléen.
  - On comparera alors cos ($V(R)$, $V(D_k)$) pour tout k dans {1,n}.
  - Classement par valeur relative (les meilleurs cos) des documents par rapport à R, ou classement avec seuil.

Comparer des documents et des requêtes

- Si les documents ont déjà été classés entre eux avec des seuils, on a une topologie boule (centre, rayon).



distance min = 2(1-σ) majorée par 1. Si V(R) est dans la boule, il y aura au plus cette distance.

(Ce qui est pertinent pour D_{k0} est aussi localement pertinent pour R.)

Mais attention : perte d'information si V(R) est en bord de boule….

Comparer des documents et des requêtes

- La meilleure solution est de créer une boule de centre V(R) et de rayon 1-σ.
  - Ne pas utiliser des classements déjà faits à cause de la non transitivité. Les distances s'accumulent.
- Quant aux cliques, elles sont plus difficiles à utiliser telles quelles mais il existe des théories les mettant en œuvre : les graphes de Gabriel.

Le modèle fréquentiel simple

- Conclusion sur le modèle booléen :
  - Les mêmes défauts que sur le document unique
  - Atténués si les vecteurs de la base sont fréquentiels.
- Proposition de plusieurs choix de représentation fréquentielle pour les documents.

Le modèle fréquentiel simple

- Valeur de la ième composante a_i^k du vecteur de document $V(D_k)$:
  - $f^k(t_i)$: fréquence d'occurrence du terme t_i de B dans le kième document
  - $f^k(t_i) / \| V(D_k) \|$: vecteur normé
  - $f^k(t_i) \times p_i$ avec ou sans $1/ \| V(D_k) \|$: fréquence d'occurrence avec pondération de la catégorie grammaticale du terme t_i.
- Modèle identique pour le vecteur de requête.

Le modèle fréquentiel simple

- Plus le nombre de documents dans la base est grand, plus il y a de chances que B soit grand.
- L'union des $V(D_k)$ risque de composer une matrice très creuse.
- On peut proposer les mêmes solutions de réduction que celles présentées pour le document unique
- Elles sont plus raisonnables dans ce cas que dans celui du document unique.
  - Un terme qui n'apparaît qu'une seule fois dans 100 documents différents a peu de chances d'être représentatif.

Le modèle fréquentiel simple

- Toutes les techniques :
  - Élimination des fréquences faibles avec ou sans pondération préalable des catégories
  - Élimination des mots de bonne catégorie mais d'usage trop courant
- Peuvent être utilisées.
- Défauts :
  - Possibilité de plus grand silence pour les requêtes, mais il y a un compromis à trouver.
  - Ce compromis, Salton propose de le rendre dépendant de la notion de **valeur discriminatoire** d'un mot par rapport à un document.

Le modèle tf/idf

- Term Frequency / Indexing Document Frequency.
- Dans la base d ’un document D_k, on calcule la fréquence d ’occurrence tf_{ki} d ’un terme t_i .
- La pondération fournie par Salton (1990)pour désigner l ’importance **discriminatoire et sémantique** de ce terme est la suivante :
  - $W_{ki} = (\log(tf_{ki}) + 1)/ \sqrt{(\Sigma_{k=1}^{N}(\log(tf_{ki}) +1))}$
  - N est le nombre total de documents considérés.
  - Quand il s ’agit d ’un document unique le poids est de 1.

Le modèle tf/idf

- W_{ki} est la valeur de la composante a_i^k du vecteur de document $V(D_k)$ sur le terme t_i.
- On remarquera que ces pondérations font les valeurs des composantes de vecteurs de documents ne sont pas réellement indépendantes entre elles.
  - L'idée que les vecteurs de documents recomposent une autre base vectorielle est mise de côté.
- Elles représentent l'idée de :
  - L'importance d'un mot par rapport à un document, comparé aux autres documents.

Traitement des requêtes

- Soit D_j un document sur lequel on veut faire une requête Q .
- Dans l 'espace $\boldsymbol{B}$, D_j est représenté par un vecteur $D_j = (d_{j1}, d_{j2},\ldots,d_{jn})$ où chacun des éléments correspond au poids du terme t_{iB} de la base B (et donc le poids de la composante vectorielle V_{iB} dans $\boldsymbol{B}$). Ce poids est calculé selon la formule précédente (en fonction de la fréquence d 'occurrence du terme t_{iB} dans D_j).

Traitement des requêtes

- De la même manière on considère que Q est représenté dans l ’espace $\boldsymbol{B}$ par un vecteur :
  - $\vec{Q}= (q_1,q_2,\ldots q_n)$ où q_i est le poids du terme t_{iB} (de la base) dans la requête Q.
    - Ce poids est calculé aussi en tf/idf par rapport aux documents de la collection.
- L ’appariement de la requête et du document se fait par le calcul de la similarité entre les vecteurs Q et D_j qui est donnée par la formule du cosinus.

Le modèle tf/idf

- Avantages :
  - Plus de précision que les modèles fréquentiels simples.
  - La base ne pourra être rétrécie que pour les termes rares dans tous les documents.
  - On peut faire des sous-collections de documents, indexés par les termes dont la pondération est la plus forte (classification « sémantique » des documents).
  - On a moins besoin d'une ontologie du domaine
- Plusieurs autres problèmes non résolus, intrinsèques au modèle de Salton.

Particularités du modèle vectoriel de Salton

- Une base par « collection de documents »
  - La dimension peut varier
  - Les composantes peuvent varier
  - Base vraie
- Représentation d ’un document par un vecteur unique
  - Les poids des termes varient en fonction de l ’état de la collection de documents => à recalculer entièrement pour tout nouveau document

Limites du modèle de Salton : aspects « syntaxiques »

- Création de la base et des vecteurs par des méthodes manuelles ou semi-automatiques au mieux.
- Dépend fortement de la qualité de la lemmatisation.
- Ne gère pas les ambiguïtés de lemmatisation (affectation de multiples catégories grammaticales). Erreurs possibles.
- Efface complètement la structuration syntaxique qui définit le gouvernement.
- Perd de vue la non commutativité du langage naturel.
  - Calcul du sens / sens du calcul
  - Médecins de ville / ville de médecins

Limites du modèle de Salton

- Les termes qui ne sont pas dans la base ne sont pas représentés
  - Silence sur les synonymes
- Les mesures permettent de constater la pertinence d ’un document par rapport à une requête et donc il s ’agit d ’une mesure de **pertinence** pas d ’une mesure de proximité thématique.
  - Problème de la **polysémie** : le terme « campagne » veut dire à la fois lieu rural, mais aussi processus d’adhésion populaire (campagne électorale) ou de combat (campagne militaire).

Quelques extensions

- Le problème des groupes nominaux prépositionnels : Formes N1 de N2 « figées » .
  - Exemple : « indice de confiance » est différent de « indice » U « confiance »
- Le problème des groupes adjectivaux « techniques » où le sens est porté par la catégorie la plus faible.
  - Exemple : boîte crannienne, glandes endocrines.
- Introduction des co-occurrents ou bigrammes :
  - Information mutuelle de Church P(A,B) = P(A)xP(B) / Log(p(A))xlog(p(B)).
- Constitutions de paires
- Constitutions de dictionnaires de substituents.

Représentation des requêtes

- Existence possible de connecteurs :
  - Négation : non traitée
  - Et
  - Ou
- Omission ou symboles ensemblistes de remplacement.
- Sémantique des connecteurs en Français :
  - Le et se comporte comme un ou logique (à cause de l'ellipse), avec des dominante soit vers le « et » soit vers le « ou ».
  - Le ou se comporte comme le ou exclusif logique.

Modèle de Salton en langage naturel

- Représentations de la sémantique du langage naturel :
  - La non représentation de tous les termes de la langue est un problème.
    - => Une base avec les 70000 mots d ’un dictionnaire ?
    - Sinon que choisit-on comme base ?
  - Le nombre de productions en langue (discours) est infini : N ->∞. Comment calcule t-on les poids ? De plus, il est inaccessible.

Modèle de Salton en langage naturel

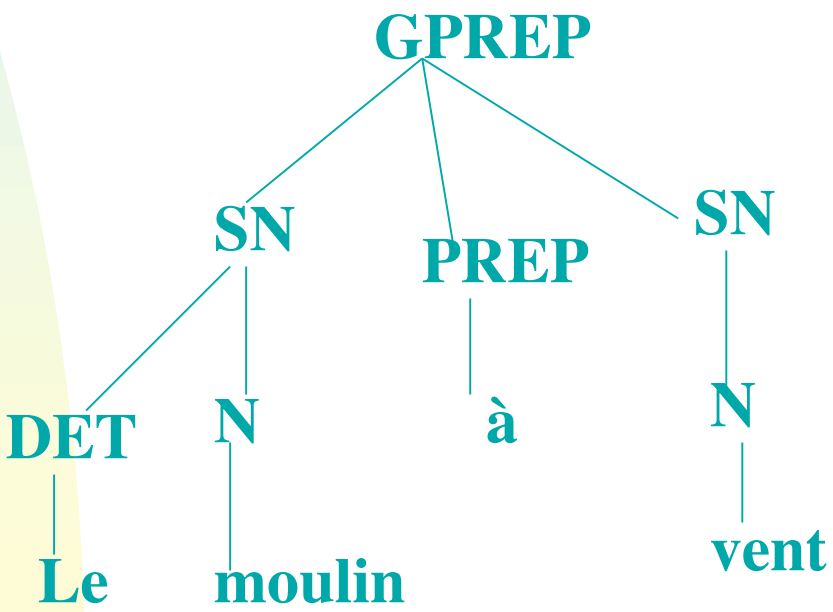
- Représentations de la sémantique du langage naturel :
  - La pertinence par co-occurrence de termes n ’est pas le problème de la sémantique:proximité thématique entre t_{iB} et $t_{(i+1)B}$? .
  - Une requête est appariée avec un document contenant les mêmes termes qu ’elle. Comment l ’apparier avec un document comprenant des termes synonymes ? => fonctions lexicales.

Modèle de Salton en langage naturel

- La syntaxe et la sémantique en langage naturel ne sont pas indépendantes.
  - La voile du bateau et le bateau à voile donnent la même requête Q={voile, bateau}= {bateau, voile}
  - Les fonctions syntaxiques analytiques donnent des informations importantes sur le rôle sémantique (casuel) des portions de textes.
    - Un sujet et un complément de manière n ’ont pas le même poids dans un texte.

Exemple

- **Sur un groupe nominal prépositionnel N1 [prep] N2.**
  - Le moulin à vent.
  - Analyse syntaxique :



Des vecteurs pour les constituents syntaxiques

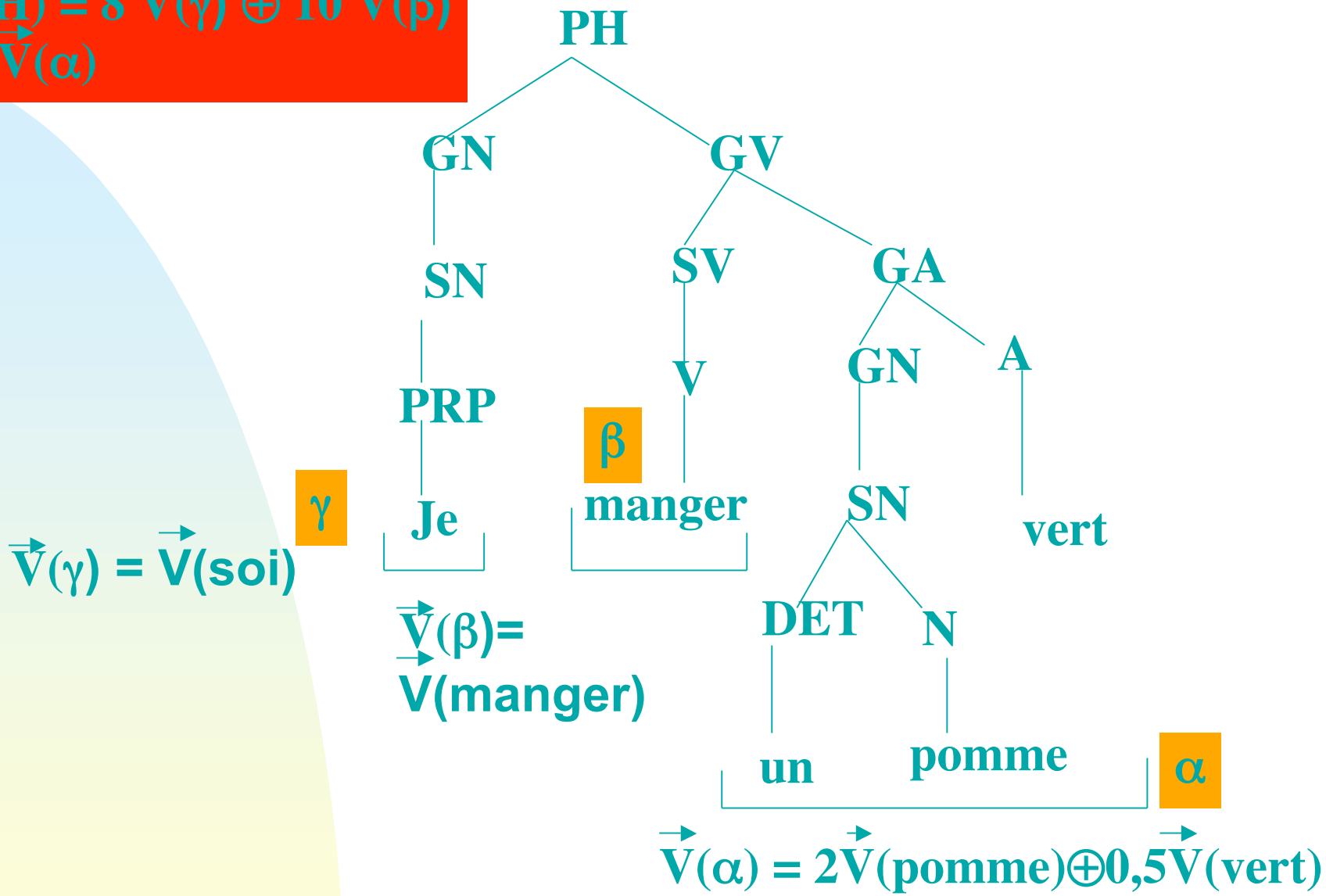
- La structure syntaxique donne:
  - Un rôle de gouverneur pour « moulin » => poids de moulin dominant, par ex. 2
  - La structure (det, N,« à », N) indique un rôle de complément circonstanciel pour le deuxième élément. => poids de « vent » gouverné, par ex. 0, 5.
- On devrait avoir un vecteur pour « moulin à vent » de la forme:
  - V(groupe(moulin à vent)) = 2V(moulin)⊕0,5 V (vent)

Mais aussi pour les phrases.

- Un texte n'est pas « un sac de mots ».
- Il est composé d'unités de sens qui sont des phrases.
- Ces phrases sont des compositions particulières.
- On devrait avoir des vecteurs pour les phrases qui tiennent compte de l'importance de chaque mot. .
- Exemple :

Je mange une pomme verte.

$\vec{V}(PH) = 8\,\vec{V}(\gamma) \oplus 10\,\vec{V}(\beta) \oplus 6\,\vec{V}(\alpha)$



Segments d ’ordre supérieur

- De la même manière, des ensembles structurés par des règles de présentation de documents comme les paragraphes, ou les documents eux-mêmes, doivent être vus comme des unités « thématiques ».
- La composition de ces unités n’est pas un collage de mot, mais une composition de toutes les unités sémantiques complexes qu’elles contiennent , les phrases.

Segments d ’ordre supérieur

- Le vecteur d ’un ensemble de phrases (paragraphe, texte) est le barycentre des vecteurs de phrases.
  - Si T= {PH_1, PH_2, …, PH_n}
  - Alors V(T) = V(PH_1) ⊕V(PH_2) ⊕ … ⊕ V(PH_n)
- De la même manière, si D est un ensemble de textes :
  - D= {T1, T2, …, Tm} alors V(D) = V(T1) ⊕V(T2) ⊕… ⊕ V(tm)

Effets de macro-structuration

- On peut, dans un texte, ou dans un ensemble de textes, tenir compte d ’un effet « d ’accroche » sémantique (ou non) d ’un sous-ensemble par rapport à un autre en substituant au vecteur barycentre un vecteur moyen pondéré.
- Exemple: l ’introduction d ’un article est un sous-texte pour lequel on peut estimer que son vecteur « pèse plus » que celui d ’un sous-texte quelconque du corps de l ’article. => **catégorisation** d ’un genre donné de documents.

Nature de la requête

- Une vraie requête en LN est :
  - Une phrase (unité sémantique complexe)
  - Un texte (unité thématique)
- Apparier une requête et un ou plusieurs textes doit se faire en considérant ces aspects.
- Quelle que soit la mesure de similarité utilisée, cosinus, ou autre, il ne faut pas perdre des informations.