Vers une détection en temps réel de document Web centrés sur une entité

Ludovic Bonnefoy*, Vincent Bouvier**, Romain Deveaud* et Patrice Bellot**

* LIA – Université d'Avignon

** LSIS – Université d'Aix-Marseille

Introduction

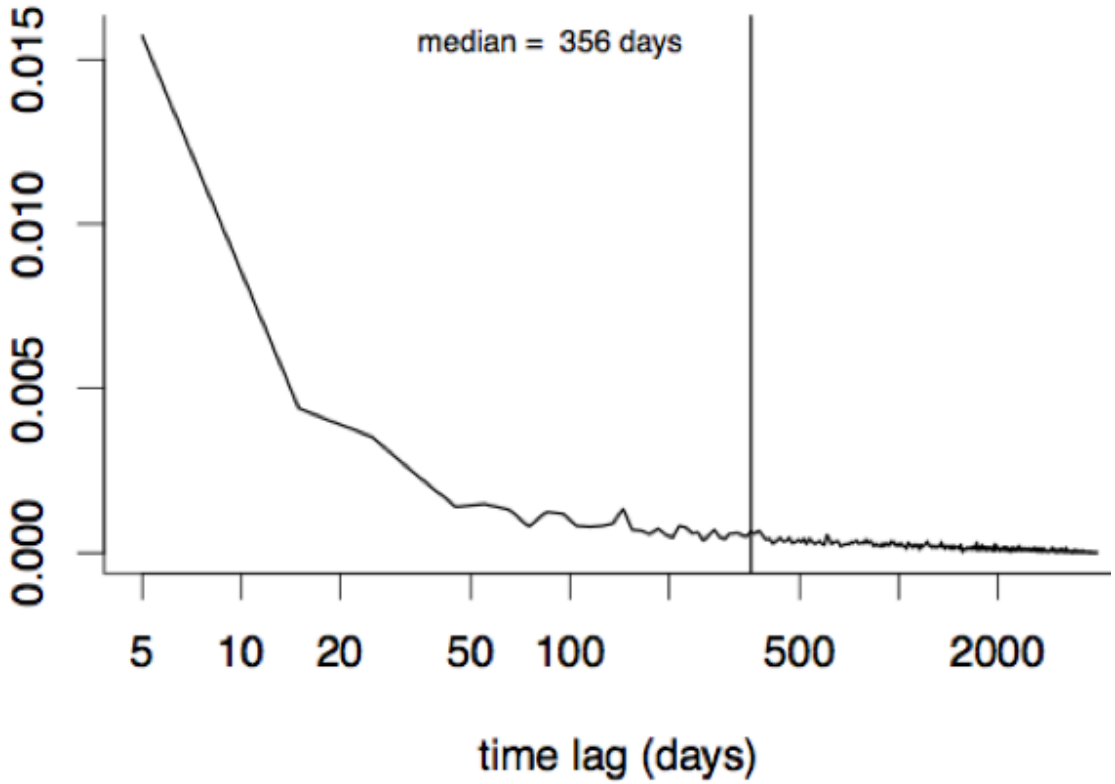
La plupart des « connaissances » de l'humanité est accessible à tous via Internet.

Du moins à ceux qui savent bien chercher..

Sinon, il existe des bases de connaissances généralistes : Wikipédia, Freebase, etc.

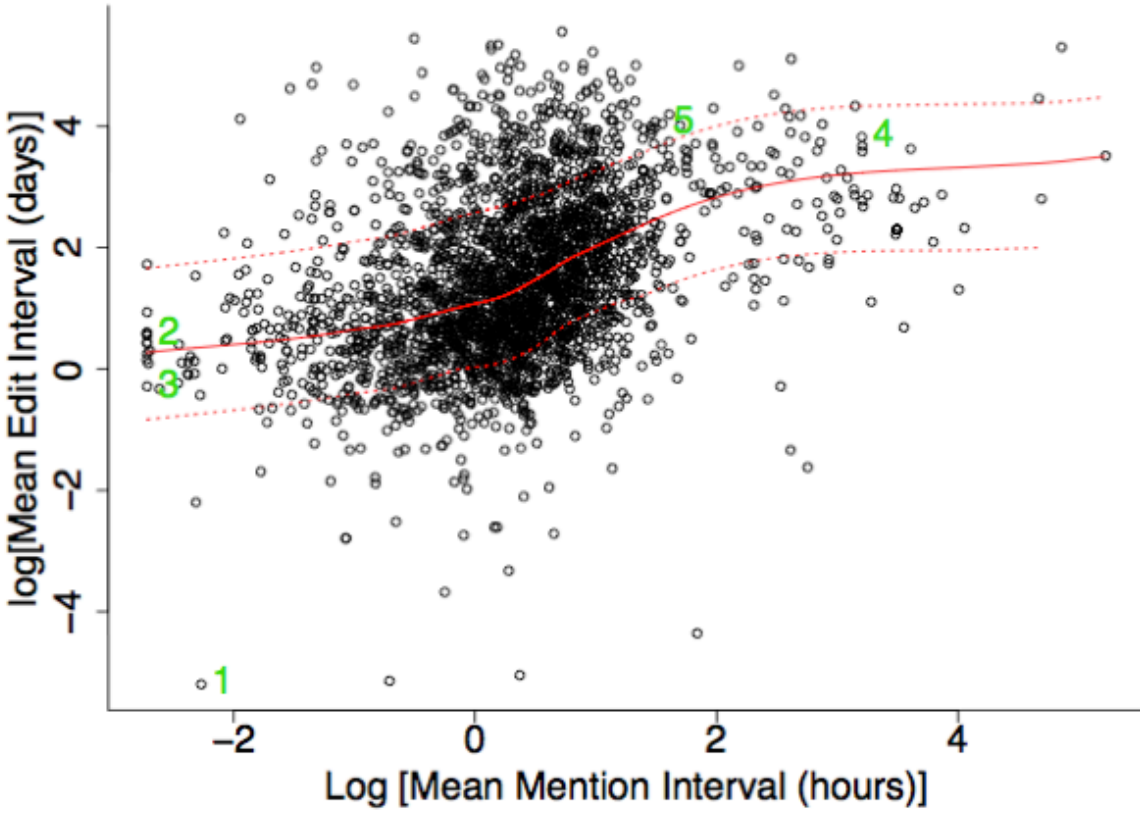
Mais sont-elles à jour ?

#contributeurs << #sujets << #documents sur ces sujets



Nombre de jours entre l'ajout d'une source sur Wikipédia et sa date de création (Frank, 2012).

Echantillon de 60000 pages web citées dans des articles de personnes.



Popularité d'une entité vs. Nombre de jours moyen entre deux éditions de l'article Wikipédia.

Moyennes calculées sur cinq semaines de documents journalistiques.

(1) Death of Michael Jackson, (2) Muammar Gaddafi, (3) Barack Obama, (4) Aung Myint Oo, (5) Allan Asher

Google Knowledge Graph

+ de 500 millions d'objets

+ de 20 milliards de faits

- 20% des entrées obsolètes en quelques semaines
  (Etude menée par Conductor)
- dont 50% avec 2 jours de retard sur Wikipédia

Objectif

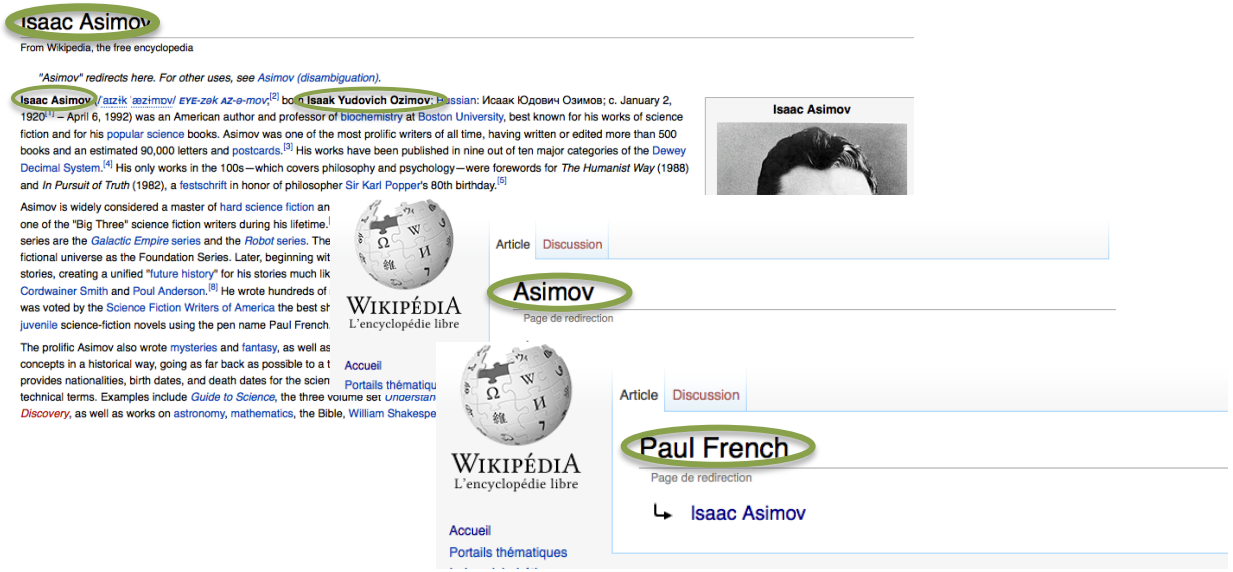
Réduire le temps de latence entre la publication sur le Web d'une information importante au sujet d'une entité et son ajout dans Wikipédia

Fournir aux contributeurs une liste de nouveaux documents susceptibles de contenir une telle information.

Une approche en trois temps

1. Sélection d'une entité, recherche de variantes d'écriture
2. Détection automatique et à la volée des documents faisant référence à l'entité
3. Déterminer si le document contient des informations pertinentes

Extraction de variantes (Cucerzan, 2007)



Nom seul	Rappel = 0,68
Nom + variantes	Rappel = 0,78

Détection et évaluation des documents

Un classifieur par décision

Mêmes critères, organisés en trois types :

- Caractéristiques intrinsèques du document
- Adéquation avec les connaissances a priori sur l'entité
- Caractéristiques de la période d'apparition du document

Caractéristiques intrinsèques du document

Un Picasso pour cent euros

Qui n'a jamais rêvé d'avoir un Picasso à la maison ? Pour 100 euros seulement, le prix d'un billet de loterie, tout le monde peut tenter sa chance. Il suffit de serendre sur le site internet, www.1picasso100euros.com. L'opération en question est une tombola en ligne organisée par l'Association internationale pour la sauvegarde de Tyr (Aist) avec le soutien de Sotheby's International dans le but de récolter des fonds pour soutenir Tyr, ville du Liban classée au patrimoine mondial de l'Unesco.

Cinquante mille billets sont mis en vente. L'heureux gagnant pourra accrocher sur l'un de ses murs un dessin de Pablo Picasso, intitulé *l'Homme au gibus*. Cette gouache sur papier mesure 30,5 cm par 24 cm, le peintre espagnol l'a réalisé en 1914, elle est estimée à 1 million de dollars (783 000 euros). Maya Picasso (fille de Pablo Picasso et Marie-Thérèse Walter), ainsi que Claude Ruiz-Picasso (fils de Pablo Picasso et Françoise Gilot) en certifient l'authenticité. Les bénéfices de la collecte seront destinés au financement d'un village d'artisanat traditionnel et à la création d'un institut d'études cananéennes, phéniciennes et puniques à Beyrouth.

L'idée du projet revient à Péri Cochin dont la mère a créé l'Aist, il y a trente ans. L'animatrice et productrice de télévision franco-libanaise a trouvé cette opération plus originale pour lever des fonds que les sempiternelles dîners de gala, explique Nathalie Zaquin, qui a participé au montage juridique de l'opération. Officiellement lancé lundi au PAD (Pavillon des arts et du design), l'initiative devrait rayonner à travers le monde, en Angleterre, aux Etats-Unis et au Moyen-Orient, grâce au soutien de la maison de vente, Sotheby's.

- p(e,d) : probabilité de l'entité dans le document
- p(e,t_d) : dans le titre
- $p_{10\%}$(e,d) : par tranche de 10%

$$p(e,x) = \frac{\text{nombre d'occurrences de e dans x}}{\text{nombre de termes dans x}}$$

Adéquation avec les connaissances a priori sur l'entité

Un Picasso pour cent euros

Qui n'a jamais rêvé d'avoir un Picasso à la maison ? Pour 100 euros seulement, le prix d'un billet de loterie, tout le monde peut tenter sa chance. Il suffit de serendre sur le site internet, www.1picasso100euros.com. L'opération en question est une tombola en ligne organisée par l'Association internationale pour la sauvegarde de Tyr (Aist) avec le soutien de Sotheby's International dans le but de récolter des fonds pour soutenir Tyr, ville du Liban classée au patrimoine mondial de l'Unesco.

Cinquante mille billets sont mis en vente. L'heureux gagnant pourra accrocher sur l'un de ses murs un dessin de Pablo Picasso, intitulé *l'Homme au gibus*. Cette gouache sur papier mesure 30,5 cm par 24 cm, le peintre espagnol l'a réalisé en 1914, elle est estimée à 1 million de dollars (783 000 euros). Maya Picasso (fille de Pablo Picasso et Marie-Thérèse Walter), ainsi que Claude Ruiz-Picasso (fils de Pablo Picasso et Françoise Gilot) en certifient l'authenticité. Les bénéfices de la collecte seront destinés au financement d'un village d'artisanat traditionnel et à la création d'un institut d'études cananéennes, phéniciennes et puniques à Beyrouth.

L'idée du projet revient à Péri Cochin dont la mère a créé l'Aist, il y a trente ans. L'animatrice et productrice de télévision franco-libanaise a trouvé cette opération plus originale pour lever des fonds que les sempiternelles dîners de gala, explique Nathalie Zaquin, qui a participé au montage juridique de l'opération. Officiellement lancé lundi au PAD (Pavillon des arts et du design), l'initiative devrait rayonner à travers le monde, en Angleterre, aux Etats-Unis et au Moyen-Orient, grâce au soutien de la maison de vente, Sotheby's.

Similarité

Cosinus

Entités liées à Picasso

Pablo Picasso

« Picasso » redirige ici. Pour les autres significations, voir Picasso (homonymie).

Pablo Ruiz Picasso, né à Malaga, Espagne, le 25 octobre 1881 et mort le 8 avril 1973 à Mougins, France, est un peintre, dessinateur et sculpteur espagnol[1] ayant passé l'essentiel de sa vie en France.

Artiste utilisant tous les supports pour son travail, il est considéré comme le fondateur du cubisme avec Georges Braque et un compagnon d'art du surréalisme. Il est l'un des plus importants artistes du XXe siècle, tant par ses apports techniques et formels que par ses prises de positions politiques. Il a produit près de 50 000 œuvres dont 1 885 tableaux, 1228 sculptures, 2880 céramiques, 7089 dessins, 342 tapisseries, 150 carnet de croquis et 30 000 estampes (gravures, lithographies, etc), [2].

Cubisme

De 1907 à 1914, il réalise avec Georges Braque des peintures qui seront appelées « cubistes ». Elles sont caractérisées par une recherche sur la géométrie et les formes représentées : tous les objets se retrouvent divisés et réduits en formes géométriques simples, souvent des carrés. Cela signifie en fait qu'un objet n'est pas représenté tel qu'il apparaît visiblement, mais par des codes correspondant à sa réalité connue. Le cubisme consiste aussi à représenter sur une toile en deux dimensions un objet de l'espace. Picasso décompose l'image en multiples facettes (ou cubes, d'où le nom de cubisme) et détruit les formes du réel pour plonger dans des figures parfois étranges (comme une figure représentée sur une moitié de face, et sur l'autre de côté). Cette technique, initiée par Picasso et Braque, fit de nombreux émules tels que Juan Gris, Francis Picabia, Brancusi, les Delaunay, Albert Gleizes.

Descendance

Picasso a eu quatre enfants :

- Paulo Picasso (4 février 1921 - 5 juin 1975), avec Olga Khokhlova
- Maya Widmaier-Picasso (née le 5 septembre 1935), avec Marie-Thérèse Walter
- Claude Picasso (né le 15 mai 1947), avec Françoise Gilot
- Paloma Picasso (née le 19 avril 1949), avec Françoise Gilot

$Sim_{1g}(d,w_e)$: unigrammes

$Sim_{2g}(d,w_e)$: bigrammes

p(el, d) : probabilité des entités liées dans d

Caractéristiques de la période d'apparition du document

- DF(e, 1j), DF(e,7j) : Nombre de documents mentionnant l'entité ce jour-ci ou les 7 derniers jours
- Variance de DF(e,7j)
- TF(e, 7j) : Nombre de mentions en 7 jours
- TF(e, titres, 7j) : Nombre de mentions dans les titres des documents en 7 jours

Choix des classifieurs

Forêt d'arbres de décisions :

- Chaque arbre est entrainé avec un sous-ensemble des critères et des exemples
- Chaque arbre vote pour une classe

Avantages :

- Performances état de l'art
- Pas de sur-apprentissage
- Estimation de l'importance de chaque critère

TREC Knowledge Base Acceleration (KBA)

Pour chacune des 29 entités évaluées, associer aux documents la mentionnant un score de pertinence (entre 0 et 1000)

Corpus :

	Presse	Web	Social
# docs	134 625 663	5 400 200	322 650 609
taille	8072GB	350GB	531GB

- Collecté sur une période de 7 mois, soit 4973 heures.
- Deux parties : du 10/11 au 12/11 pour entraînement et du 01/12 au 04/12 pour l'évaluation

Protocole d'évaluation

Document	Score	F-mesure	Seuil
1328622420-d35dd93c72a65471da1d227273de7c34	987	0,17	
1329416220-0965e0ea31da5b1c36a20c625dd74fd7	973		950
1329422280-983af7a38f869a2e84c6c1553f95ce85	918	0,15	900
1327951740-9f41cbf2f35d73e5edda14f997b3caf7	827	0,14	800
1334185920-6a5c2ac675437f1896137e2d0f62bf68	692		
1329412740-c1cb824cdabfe49ff40cb3999b63c323	666	0,25	650
1329412440-625331f102ed4548d50492a07a8e2601	563		
1327950420-5fa264f81f943a7f01b2dc752a2332c4	555	0,33	550
1334187000-7c2eeb9d4e0071bd9ee6a379298b819c	519	0,31	500
1327951260-c1afc3708bd20d3072b7dd404c2f411a	495		
1328625240-83e099a9900d73f78ea74462422f8c8f	460		
. . .		. . .	

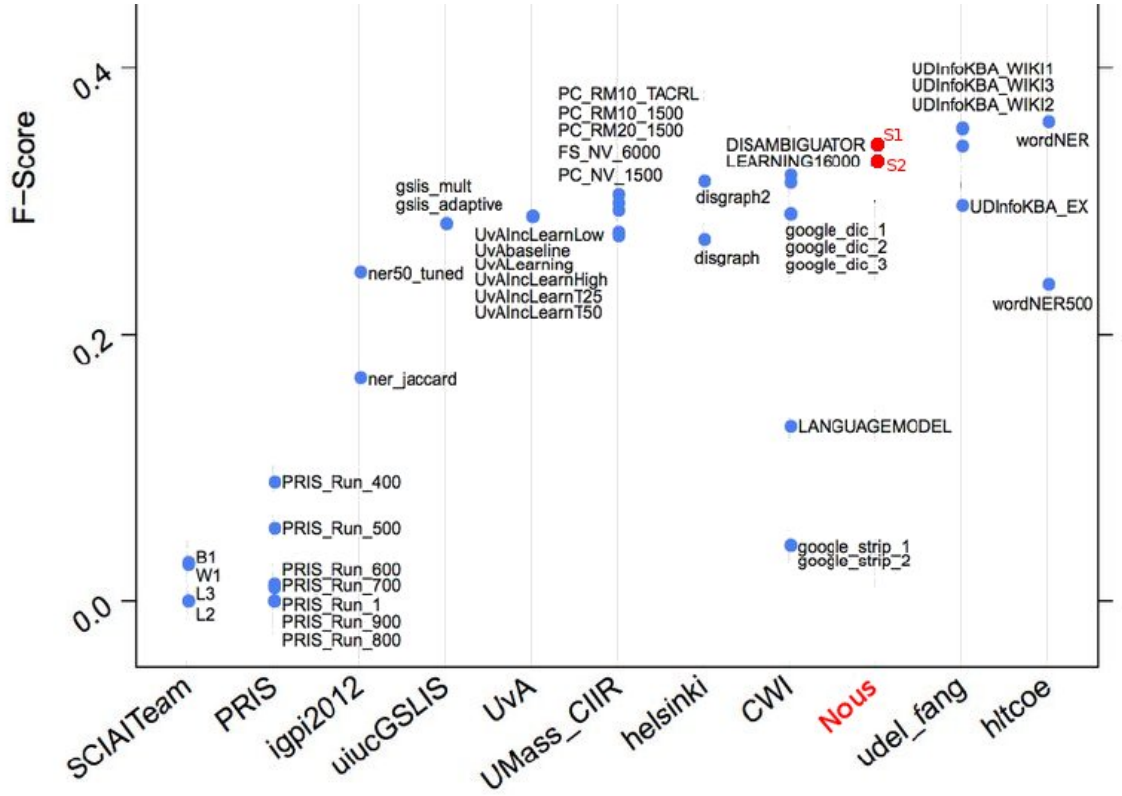
Score du système : score correspondant au meilleur seuil (moyenne des f-mesures pour chaque entité à ce seuil).

Contrairement aux autres tâches à TREC, pas de pooling, tous les documents récupérés avec une baseline sont évalués.

Adaptation de notre approche

Utilisation des scores de confiance des classifieurs pour classer les documents

$$S_1(d_i) = \begin{cases} s(d_i, c_{np,p}) \times s(d_i, c_{i,c}) \text{ si } s(d_i, c_{np,p}) >= 0,5 \\ \text{retiré de la liste sinon} \end{cases}$$



- WordNER : un classifieur svm par entité utilisant les unigrammes et entités présents dans les documents comme critères
- S1 : retrait de la liste des documents jugés non pertinents par le premier classifieur
- S2 : conservation de tous les documents

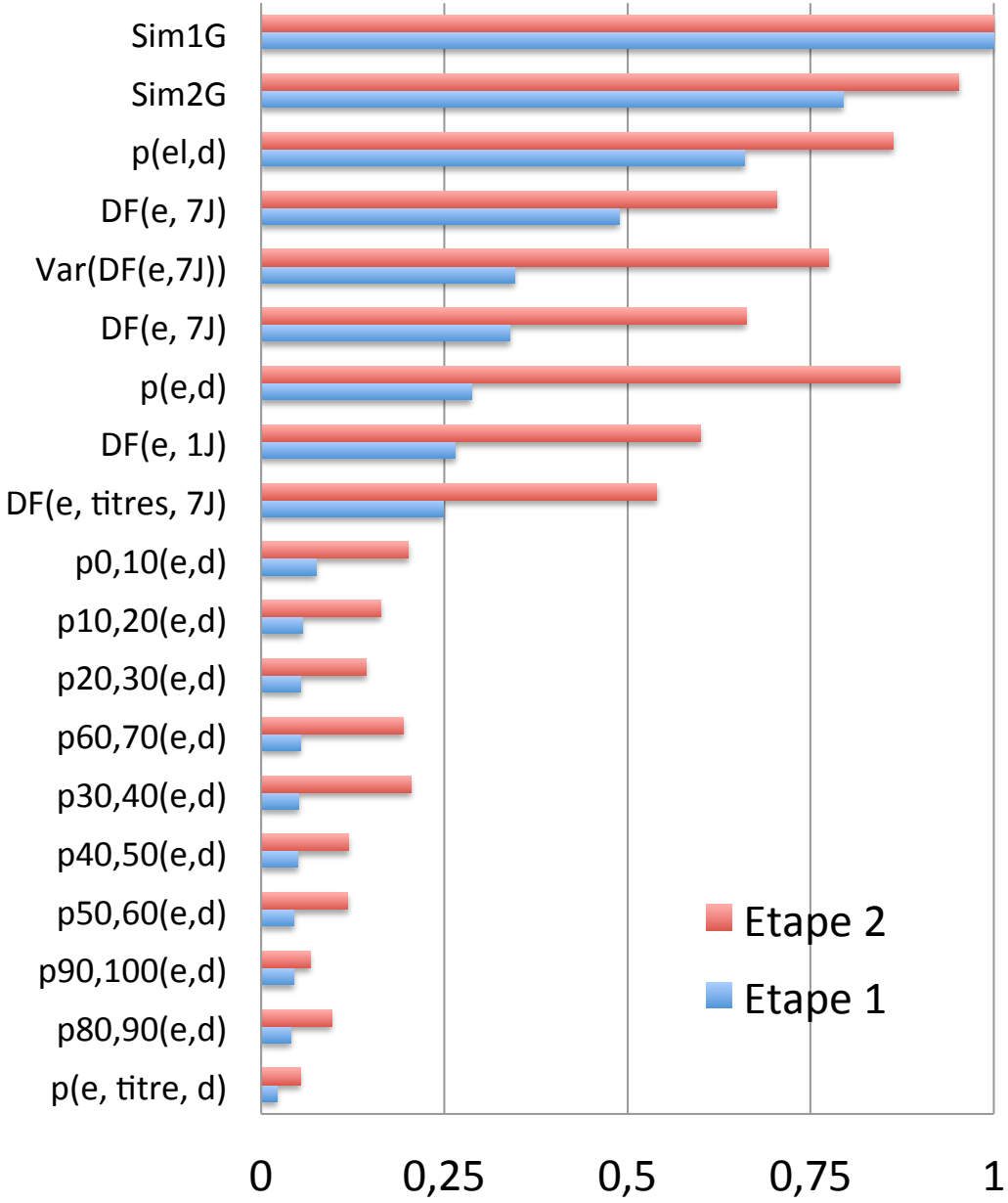
Word NER	0,359	Médiane	0,280
S1	0,342	Moyenne	0,22

Importance des critères

Coefficients de Gini ->

- Prépondérance des critères sur la cohésion avec la connaissance a priori
- Importance du facteur temps
- Position des mentions et impact du titre peu utiles

	Sans Titre	Avec Titre
Non Pertinent	48%	44%
Pertinent	52%	56%



Conclusions et perspectives

Approche en deux principales étapes, efficace pour trouver des documents contenant des informations importantes sur une entité donnée.

- La présence d'entités liées est un indice important : mettre à jour les relations de l'entité
- La similarité du document avec un document de référence est importante aussi : divergence dans le temps?
- La période d'apparition aussi : prendre en compte l'écho?

Des questions?

MERCI POUR VOTRE ATTENTION