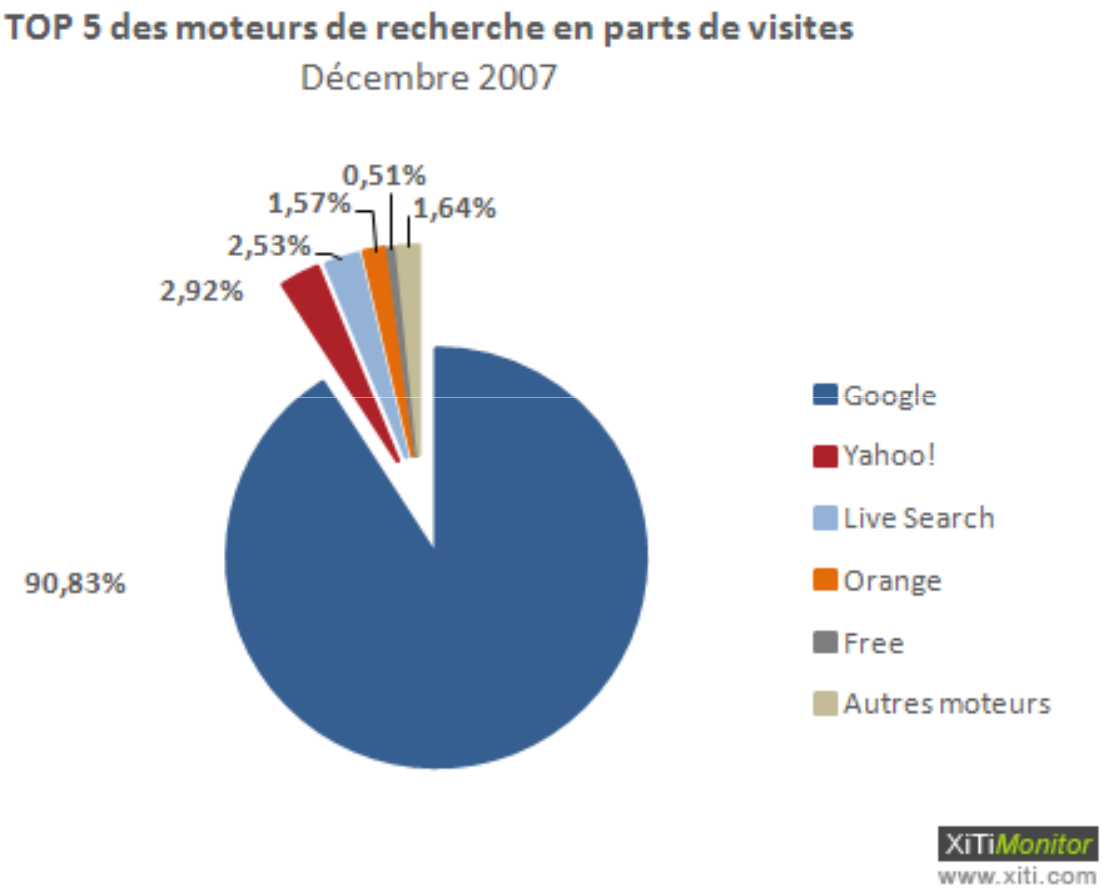
Indexation et recherche d'information en langue arabe

Ramzi ABBES, ICAR-CNRS/Lyon 2
Malek Boualem, France Télécom R&D
Mohamed Hassoun, ELICO/ENSSIB

Plan

- Moteurs de recherche ou Google
- Étude : coté requêtes
- Étude : coté corpus
- Importance des outils linguistiques
- Limites des outils linguistiques

Moteurs de recherche par défaut ou Google



Moteurs de recherche par défaut ou Google

- PageRank
- Pondération des pages +/-
  - Sensible aux comportements des internautes
  - Nombre de visites
  - Navigation dans le site
  - Les liens hypertextes
  - Priorité d'indexation
- Lemmatisation
  - Singulier – Pluriel
  - Masculin - Féminin
  - Mots sémantiquement proches « ~ »

Exploration, Indexation, Recherche

- Indépendante de la langue
  - Google, Yahoo, Exalead (COLTEC)
  - Popularité
  - المنتديات
- Moteur arabe pour la langue arabe
  - Couverture réduite
  - Durée de vie limitée
  - Utilisation locale
  - Internautes arabes = bilingues

Exploration, Indexation, Recherche

- Balise « meta »
- Balise « titre »…
- Quelques contenus
- Indexation du contenu
  - Regroupement des formes de la même famille morphologique,
    - Tailles plus ou moins importantes
    - Lemmatisation
      - كتبنا ,لكتب ,كتباً ,والكتب ,الكتب
      - أكاتب يسين
      - اللامركزية

Exploration, Indexation, Recherche

- Précisions
  - La vocalisation n’est pas prise en compte,
- Troncature
  - Essentiel à cause des clitiques
    - Proclitique : liaison, interrogation …
    - Enclitique : pronoms
  - Parcours de surface, insuffisant
    - « *قال* » E:\Fes-Freq\corpus hayet avec chadda\traitement minimalisé
    - « *كتب* »

Exploration, Indexation, Recherche

- Dissymétrie
  - كتب = كَتَب, كُتُب, كَتْب
  - شعر = شَعَر, شِعْر, شَعْر
  - علم = عَلَم, عَلِم, عِلْم
- Les noms propres
  - كاتب يسين
  - أكاتب يسين

Recherche Google

- Lemmatisation
  - « سماء », Google renvoi 594 000 pour سماء
  - 279 000 pour أسماء, الأسماء
  - أسماء السماوات
- Famille morphologique
  - كُتُب = كَتَب
  - Recherche avec كتب, donne les livres
- Y a t'il une priorité pour les noms
  - Recherche avec جمال donne …
- Et le EN
  - Pas de lien entre جمال et الزعيم

Requêtes – Corpus et répartitions

- Corpus
  - 2880 requêtes arabes
  - Sur deux mois
- Catégories grammaticales
  - 94,2% Formes nominales (تونس)
  - 3,3% formes verbales (اخترع)
  - 2,5% mots grammaticaux (من, متى)
  - كتب, نزل, طلب, شعر, رقص

Genres – Généralités

- Masculin
  - شاعر, رئيس…
- Féminin
  - Classique
    - رئيسة, أميرة
  - Avec marque et sans masculin
    - شجرة, غرفة, دراسة
  - Sans marque mais avec masculin
    - سكرى, حمراء
  - Sans marque et sans masculin
    - حبلى
  - Masculin
    - خليفة

Genres - Statistiques

- Généralité
  - ة mais pas uniquement, اثارة
- 50,13% de masculin
- 49,84% de féminin
  - 47,11% marque morphologique + masculin
  - 23,38% féminin du pluriel d’un masculin singulier
  - 16,81% Féminin sans marque, ayant un masculin
  - 11,69% Féminin avec marque, sans masculin
  - 1,01% Féminin sans marque, sans masculin

Nombre – Généralité

- Suffixation !
  - كتابة ------- كتاب + ات = كتابات
  - فتاة --------- فتات + ان = فتاتان
- Le pluriel brisé
  - رجل – رجال
  - نسوة – نساء – امرأة

Nombre - Statistiques

- 74,21% Singulier
- 1,77% Duel
- 24,02% Pluriel
  - 71,09% Pluriel brisé
  - 21,29% Pluriel féminin régulier
  - 6,19% Pluriel masculin régulier
  - 1,33% autres
- محمدين, البحرين

Noms propres

- 74,75% Pays
- 23,41% Noms/Prénoms
- 1,84% ville
- ……
- …..

Détermination

- 61,03% Indéterminé
- 37,97% Déterminé avec ال
- Augment la précision
- ..... الوان

Autres

- Origine, 100% en caractères arabes
  - 95,07% Arabe
  - 3,19% Dialecte
  - 1,74% autres
- Erreurs orthographiques, 5,73%
  - 96,36% Hamza
  - 3,64% ta marbouta
  - ….. Et au niveau du corpus

Étude sur corpus

- 4 338 articles de l'année 1995
- 2 006 631 mots arabes => 149 990 termes
  - 1 075 347, Titres de la une
  - 866 764, éditos, culture, critiques littéraires, débats, courrier des lecteurs…
  - 64 520, Automobiles
- 366 447 autres : ponctuations, chiffres, mots en caractères latins

Étude sur corpus - Pratique d'écriture - Hamza

- أ+إ= ا
- Exemple :
  - 26923 الى; 2089 إلى
  - 33901 ان; 50569 أن; 769 إن
  - Extraire des mots, hors mots outils
- Termes hors analyse; 5,79%
- Mots hors analyse; 6,76%

Étude sur corpus - Pratique d'écriture - Ya

- ى = ي
- يبقى = يبقي
- الاولى = الأولى = الاولي

Étude sur corpus - Pratique d'écriture - Ya

- ـاً => ـاً
- 1174 خصوصاً; 376 مشيراً
- 3,66% des mots,
- 2,07% des termes.

limite du TAL - classique

- Deuxième langue
  - بروكسال، بروكسيل
- Déclinaison
  - دولار: الدولارات, دولاراً, الدولار, دولارات, بالدولار, للدولار, دولارا, والدولار, ودولارهم, البترودولار, دولاران, كالدولار, ودولارات, دولارين, بدولار
  - موديل: 44 dérivés …
- Des noms propres
  - امريكا
  - كلينتن

Conclusion - Précision de la recherche

- Comment préciser la recherche La vocalisation
- Index non vocalisé
- Textes non vocalisés
- Les outils linguistiques permettront
  - D'élargir la recherche à une famille morphologique
    - Singulier – duel - pluriel
    - Masculin – féminin
    - Lemmatisation
  - Ne peut pas réduire le résultat, les textes ne sont pas vocalisés
- Nécessité de lemmatisation