THE UNIVERSITY of EDINBURGH

Search is not only the Web
IR Applications

Walid Magdy

School of Informatics
University o Edinburgh

31 Oct 2017

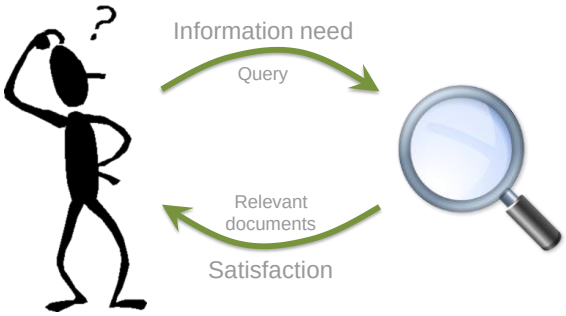
Objectives

- Main objective of IR
- Different tasks in IR
  - Printed documents search
  - Patent search
  - Social search
- This Lecture:
  High level – Simplified – Compressed

Information Retrieval Objective

- IR is finding material of an unstructured nature that satisfies an information need from within large collections.

- **Information need**
  - Expected search scenario?
  - Modeling the task?
- **Data nature**
  - Approach?
  - Scalable? Fast?
- **User Satisfaction**
  - More relevant documents?
  - Effective evaluation?

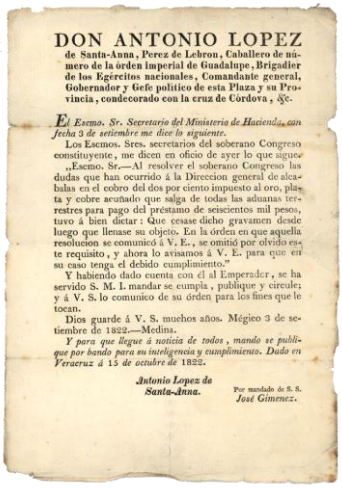


Printed Documents Retrieval

Institute for Language, Cognition and Computation
ILCC
THE UNIVERSITY of EDINBURGH

Printed Documents Retrieval

- **Documents:**
  text on printed papers (books)
- **Information need:**
  Information within these books
- **Challenge:**
  It is an image of text
- **Common Approach:**
  OCR → Recognized text ← Search
- **Challenges in Common Approach:**
  OCR → Text with mistakes ($WER_{Ar} \approx 40\%$)
  OCR → Not available for all languages



DON ANTONIO LOPEZ
de Santa-Anna, Perez de Lebron, Caballero de número de la orden imperial de Guadalupe, Brigadier de los Egércitos nacionales, Comandante general, Gobernador y Gefe politico de esta Plaza y su Provincia, condecorado con la cruz de Córdova, &c.

El Escmo. Sr. Secretario del Ministerio de Hacienda, con fecha 3 de setiembre me dice lo siguiente.

Los Escmos. Sres. secretarios del soberano Congreso constituyente, me dicen en oficio de ayer lo que sigue: „Escmo. Sr.—Al resolver el soberano Congreso las dudas que han ocurrido á la Direccion general de alcabalas en el cobro del dos por ciento impuesto al oro, plata y cobre acuñado que salga de todas las aduanas terrestres para pago del préstamo de seiscientos mil pesos, tuvo á bien dictar: Que cesase dicho gravamen desde luego que llenase su objeto. En la órden en que aquella resolucion se comunicó á V. E., se omitió por olvido este requisito, y ahora lo avisamos á V. E. para que en su caso tenga el debido cumplimiento."

Y habiendo dado cuenta con él al Emperador, se ha servido S. M. I. mandar se cumpla, publique y circule; y á V. S. lo comunico de su órden para los fines que le tocan.

Dios guarde á V. S. muchos años. Mégico 3 de setiembre de 1822.—Medina.

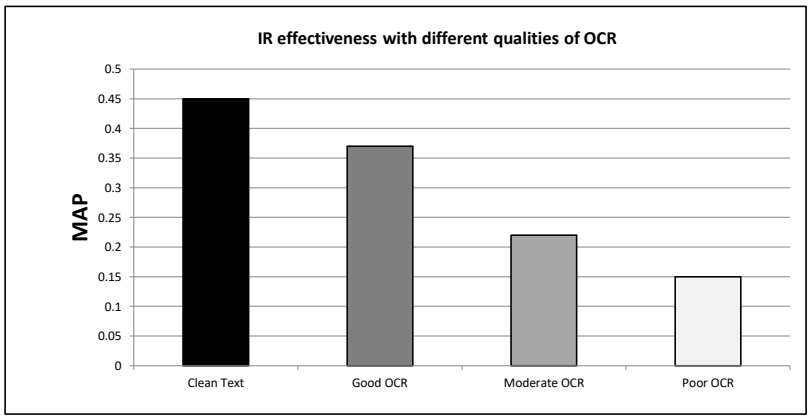
Y para que llegue á noticia de todos, mando se publique por bando para su inteligencia y cumplimiento. Dado en Veracruz á 15 de octubre de 1822.

Antonio Lopez de Santa-Anna.

Por mandado de S. S.
José Gimenez.

Problem

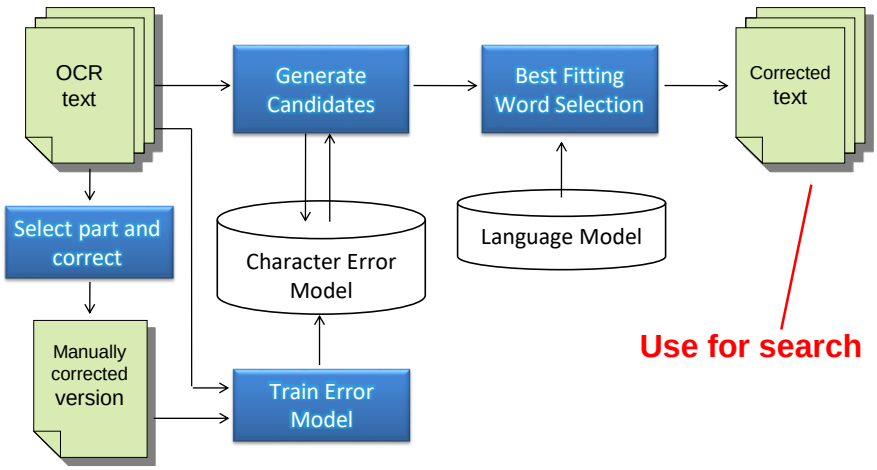
- Text with errors (sometime many errors)



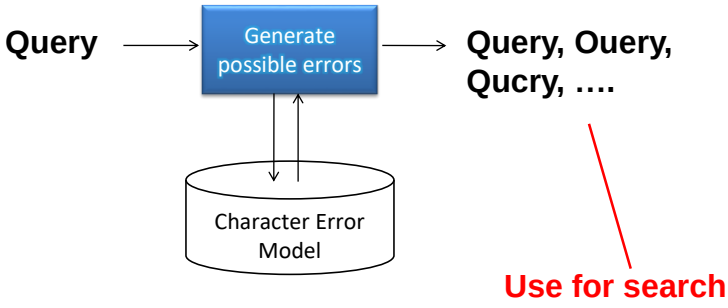
n-gram Char Representation of OCR

- **Original:**
  example sentence
- **OCR output:**
  exarnple senlcnce
- **3-gram char representation:**
  $ex exa xar arn rnp npl ple le$ $se sen enl nlc lcn cnc nce ce$
- **Query:**
  example sentence →
  $ex exa xam amp mpl ple le$ $se sen ent nte ten enc nce ce$
- **Matching:**
  $ex exa xar arn rnp npl ple le$ $se sen enl nlc lcn cnc nce ce$
  $ex exa xam amp mpl ple le$ $se sen ent nte ten enc nce ce$

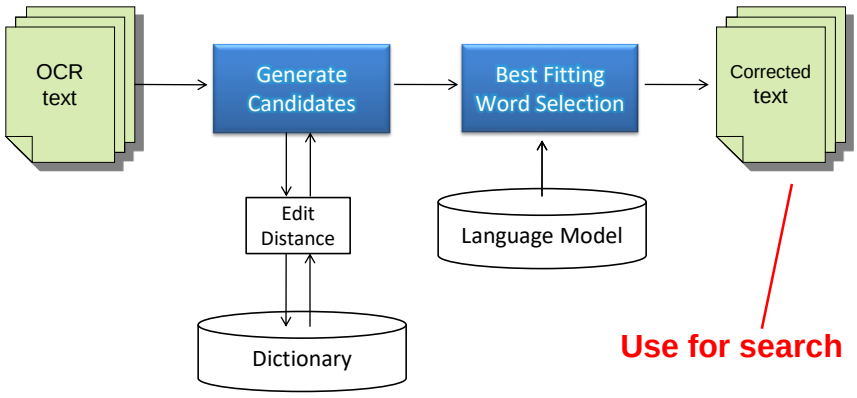
OCR Correction using Error Model



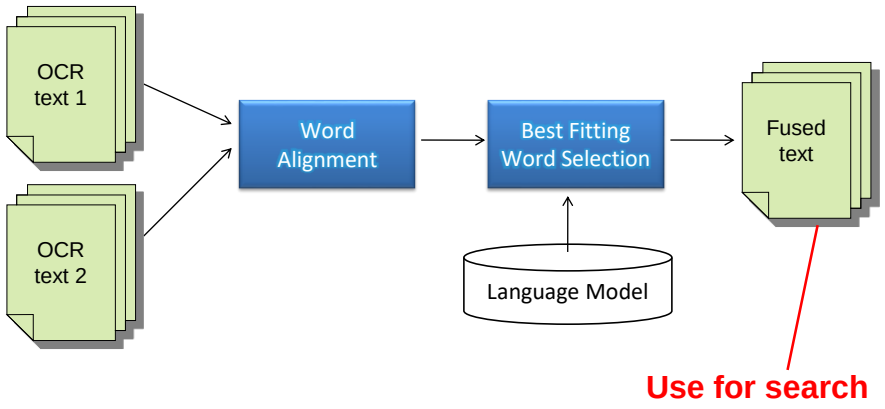
Query Garbling using Error Model



OCR Correction using Edit Distance



Multi-OCR Text Fusion

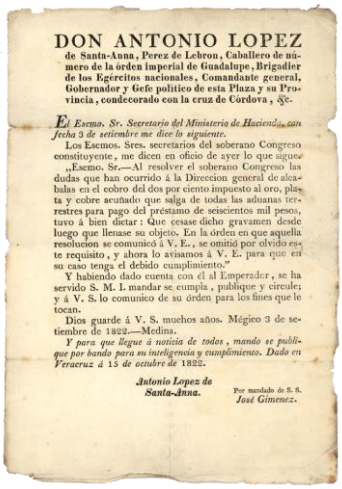


OCR Search

- Recognition errors in OCR text degrades retrieval
- Different methods of text processing can overcome the negative effect on retrieval and improves search
- n-gram character representation improves retrieval, but not that much
- Some training and resources are needed which can be manual correction, trained language model, or both
- Previous methods fail when errors are large (WER>50%)

Solution – back to Information Need

- **Information need:**
  the printed papers
- **Question:**
  Why convert image to text?
- **Related work:**
  Word Spotting



DON ANTONIO LOPEZ
de Santa-Anna, Perez de Lebron, Caballero de número de la orden imperial de Guadalupe, Brigadier de los Egércitos nacionales, Comandante general, Gobernador y Gefe politico de esta Plaza y su Provincia, condecorado con la cruz de Córdova, &c.

El Escmo. Sr. Secretario del Ministerio de Hacienda, con fecha 3 de setiembre me dice lo siguiente.

Los Escmos. Sres. secretarios del soberano Congreso constituyente, me dicen en oficio de ayer lo que sigue: „Escmo. Sr.—Al resolver el soberano Congreso las dudas que han ocurrido á la Direccion general de alcabalas en el cobro del dos por ciento impuesto al oro, plata y cobre acuñado que salga de todas las aduanas terrestres para pago del préstamo de seiscientos mil pesos, tuvo á bien dictar: Que cesase dicho gravamen desde luego que llenase su objeto. En la órden en que aquella resolucion se comunicó á V. E., se omitió por olvido este requisito, y ahora lo avisamos á V. E. para que en su caso tenga el debido cumplimiento."

Y habiendo dado cuenta con él al Emperador, se ha servido S. M. I. mandar se cumpla, publique y circule; y á V. S. lo comunico de su órden para los fines que le tocan.

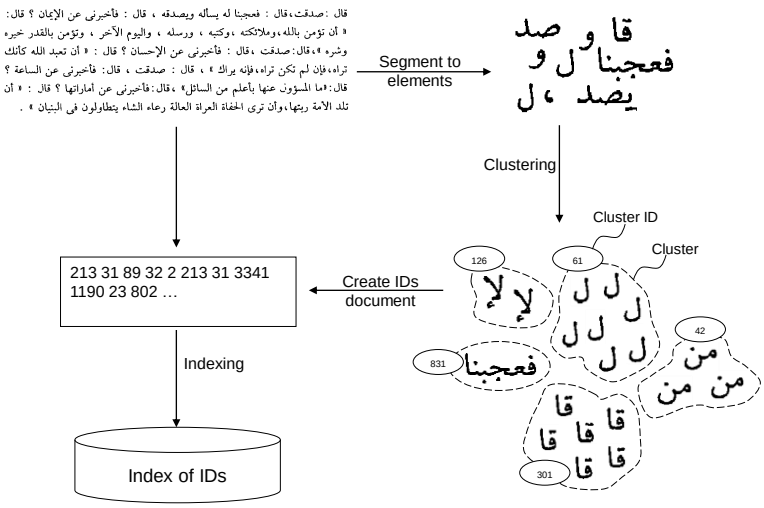
Dios guarde á V. S. muchos años. Mégico 3 de setiembre de 1822.—Medina.

Y para que llegue á noticia de todos, mando se publique por bando para su inteligencia y cumplimiento. Dado en Veracruz á 15 de octubre de 1822.

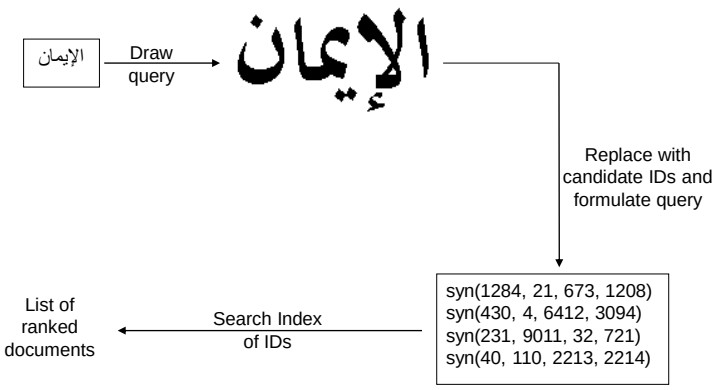
Antonio Lopez de Santa-Anna.

Por mandado de S. S.
José Gimenez.

OCRless Search



Solution – OCRless Search



Solution – OCRless Search

- **Effective and fast**
- **Robust to OCR errors** (*v1de0*)
- **No training resources required**
- **Language independent**

חֲמִישׁ العربية

English 熊 貓

- **Microsoft TechFest Demo**
  The same engine for searching printed documents in:
  Arabic, English, Chinese, Hebrew, and Hieroglyphic

Printed Documents Retrieval

- Text-based solutions: correction
- Image-based: clustering
- Current State-of-the-art: CAPTCHA

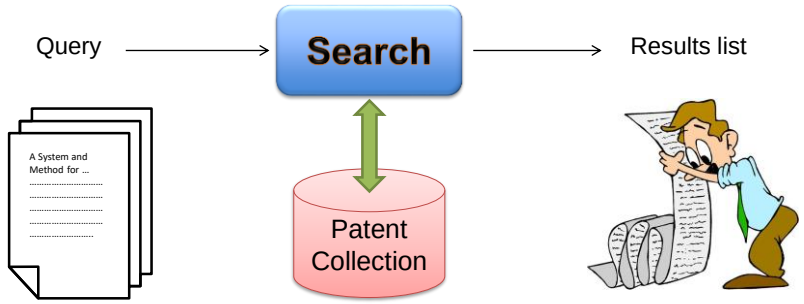
- Information need → Approach

Patent Search

Institute for Language, Cognition and Computation
ILCC
THE UNIVERSITY of EDINBURGH

Patent Search

- **Given a patent application, check if the invention described is novel**



Patent application

Several languages

Many results to check 100-600 docs/search

Patent Search – User Satisfaction

- NTCIR, CLEF, TREC
- Recall-oriented → Try not to miss a relevant document
  - Recall is the objective
- Precision is also important
- Huge # documents checked (100-600 documents)
- Evaluation: average precision (AP)!!
  - Focuses on finding relevant docs early in ranked list
  - Less focus on recall

Example

For a topic with 4 relevant docs and 1st 100 docs to be examined:

System1: relevant ranks = {1}

System2: relevant ranks = {50, 51, 53, 54}

System3: relevant ranks = {1, 2, 3, 4}

$AP_{system1} = 0.25$ $R_{system1} = 0.25$

$AP_{system2} = 0.0481$ $R_{system2} = 1$

$AP_{system3} = 1$ $R_{system3} = 1$

- We need a metric that reflects recall and ranking quality in one measure

PRES: Patent Retrieval Evaluation Score

$$PRES = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{max}}$$

n: number of relevant docs

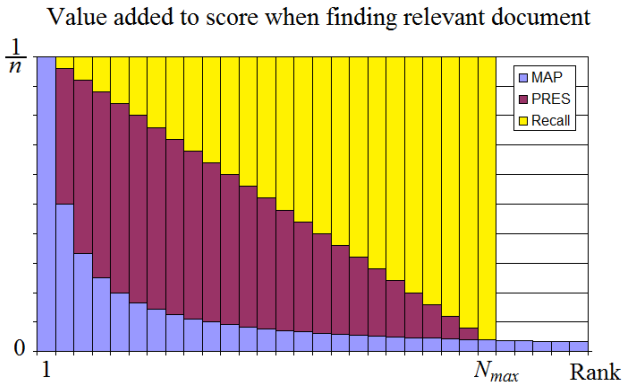
r_i: rank of the i^{th} relevant document

N_{max}: max number of checked docs

- **Derived from R_{norm} (Rocchio, 1964)**
- **Gives higher score for systems achieving higher recall and better average relative ranking**
- **Dependent on user's potential/effort (N_{max})**
- **Robust to incomplete relevance judgements**

PRES: as a cumulative gain

- **Official score in CLEF-IP since 2010**
- **Adapted in many Recall-oriented IR tasks**
- User Satisfaction → Objective function



Patent Search – CLIR

- **Query: Full patent application**
- **Common approach: MT** (the best)
- **Challenge: training recourses + speed!**
- **Ideal: Query + Document translation**

Patent Search – CLIR – Objective?

- **Manual translation**

  It is a great idea to apply stemming in information retrieval

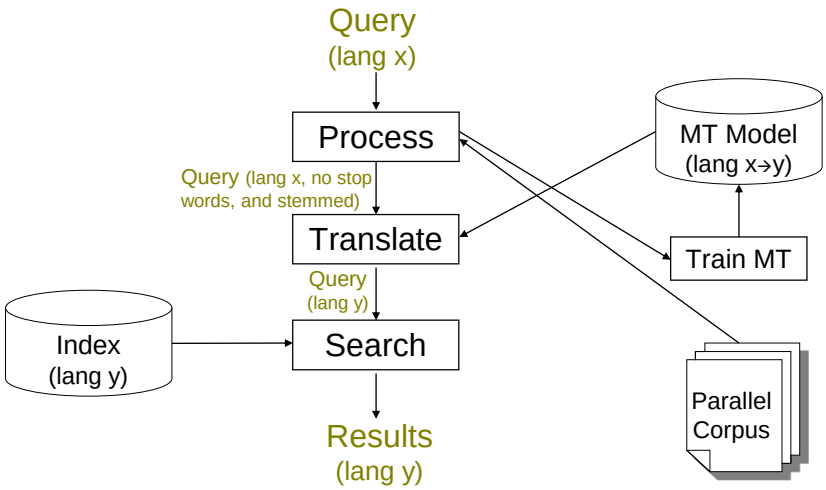
- **MT output**

he are an great ideas to applied stem by information retrieving

- **MT evaluation: MT sucks**
- **IR evaluation: MT rocks ☺**

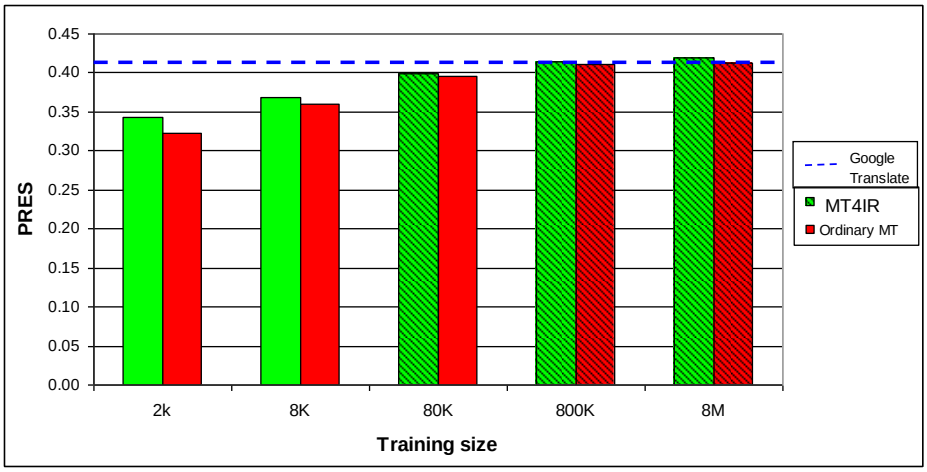
- **MT4IR:** An efficient MT that neglects morphological and syntactic features of output

Ordinary MT vs. MT4IR

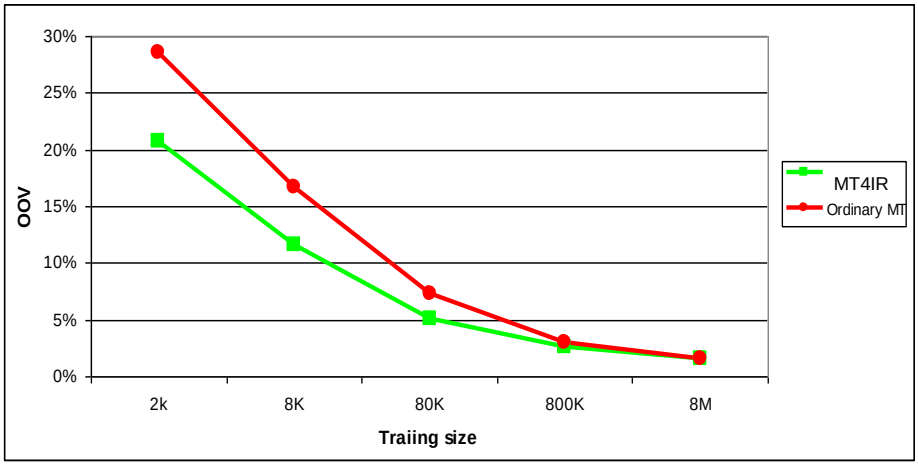


Patent Search – MT4IR

Retrieval effectiveness for a Patent CLIR **En-Fr** task



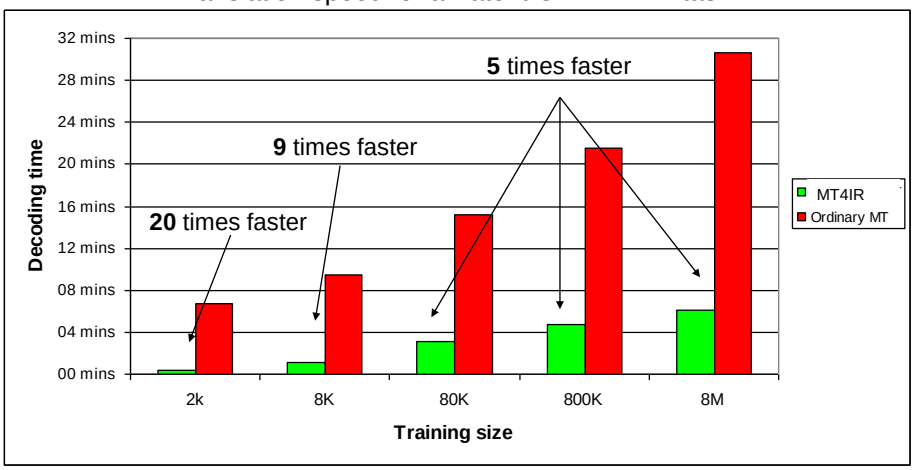
Patent Search – MT4IR



E.g. play, plays, played, playing

Patent Search – MT4IR

Translation speed for a Patent CLIR **En-Fr** task



Social Search

Microblog (Twitter) Search

Institute for Language, Cognition and Computation
ILCC
THE UNIVERSITY of EDINBURGH

Social Search

- TREC Microblog track → Ad-hoc, filtering

- User's information need?
- Search scenario? Task?
- Boolean? News updates?

Social Media & News

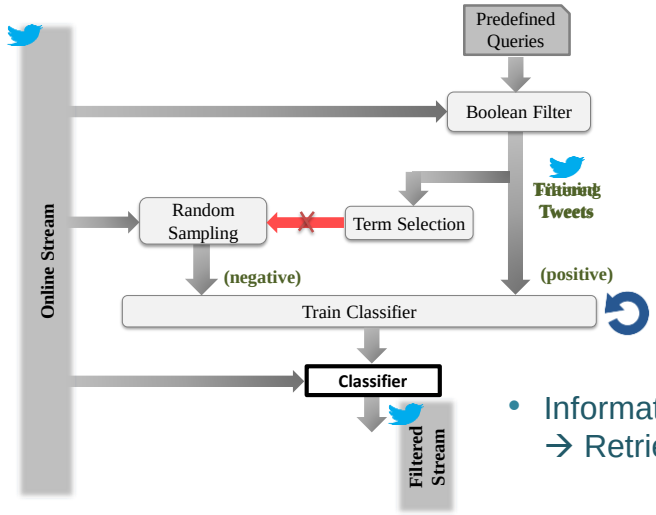
- News websites are biased

CNN vs. the guardian vs. FOX NEWS Telegraph

- People use social media to
  - Report news
  - Comment on news
  - Discuss different views on events
- Discussions on social media, reflects public interest

- Can social media answer the question:
  "*What is happening in <region>?*"

Adaptive Information Filtering



- Information need → Retrieval task

Summary

- The objective is IR is "User Satisfaction"
- Understand the user needs well
- Design the IR task carefully
- You do not have to stick to the path in the literature
- Are you sure performance is measured correctly?

- Beating the baseline is always desirable, just be sure you are moving in the right direction

Readings

- Magdy W. and T. Elsayed. Unsupervised Adaptive Microblog Filtering for Broad Dynamic Topics. *IP&M 2016*
- Magdy W. and G. J. F. Jones. Studying Machine Translation Technologies for Large-Data CLIR Tasks: A Patent Prior-Art Search Case Study. *Springer, Information Retrieval, 2013*
- Magdy W. and G. J. F. Jones. PRES: A Score Metric for Evaluating Recall-Oriented Information Retrieval Applications. *SIGIR 2010*
- Magdy W. , K. Darwish, and M. El-Saban. Efficient Language-Independent Retrieval of Printed Documents without OCR. *SPIRE 2009*
- Magdy W. and K. Darwish. Effect of OCR Error Correction on Arabic Retrieval. *Springer, Information Retrieval, 2008*