

# INTRODUCTION TO IE AND ANNIE

---

Ye Jiang and Mehmet Bakir

The Session will be recorded. The records will be available after the training



# ABOUT THIS TUTORIAL

This part of the tutorial comprises the following topics:

1. Introduction to Information Extraction (IE)
2. A Closer look at ANNIE
3. Evaluation and Corpus Quality Assurance

# INFORMATION EXTRACTION (IE)

**Information extraction (IE)** is the task of automatically **extracting structured information** from unstructured and/or semi-structured machine-readable documents and other electronically represented sources\*.

[https://en.wikipedia.org/wiki/Information\\_extraction](https://en.wikipedia.org/wiki/Information_extraction)

# NAMED ENTITY RECOGNITION (NER): THE CORNERSTONE OF IE

Traditionally, **NER is the identification of proper names in texts**, and their classification into a set of predefined categories of interest

- Person
- Organisation (companies, government organisations, committees, etc.)
- Location (cities, countries, rivers, etc.)
- Date and time expressions

Various other types are frequently added, as appropriate to the application, e.g., newspapers, ships, monetary amounts, percentages.

# WHY IS NE IMPORTANT?

- **NE provides a foundation** from which to build more complex IE systems
- Relations between NEs can provide tracking, ontological information and scenario building
  - Tracking (co-reference): “Dr Smith”, “John Smith”, “John”, “he”
  - Ontologies: “Athens, Georgia” vs “Athens, Greece”

# TYPICAL NE PIPELINE



**Pre-processing** (tokenisation, sentence splitting, morphological analysis, POS tagging)



**Entity finding** (gazetteer lookup, NE grammars)



**Coreference** (alias finding, orthographic coreference etc.)



**Exporting** to database / XML / ontology

# EXAMPLE OF IE - UNSTRUCTURED TEXT

John lives in London. He works there for Polar Bear Design.

# EXAMPLE OF IE - TOKENIZE

John lives in London . He works there for Polar Bear Design .

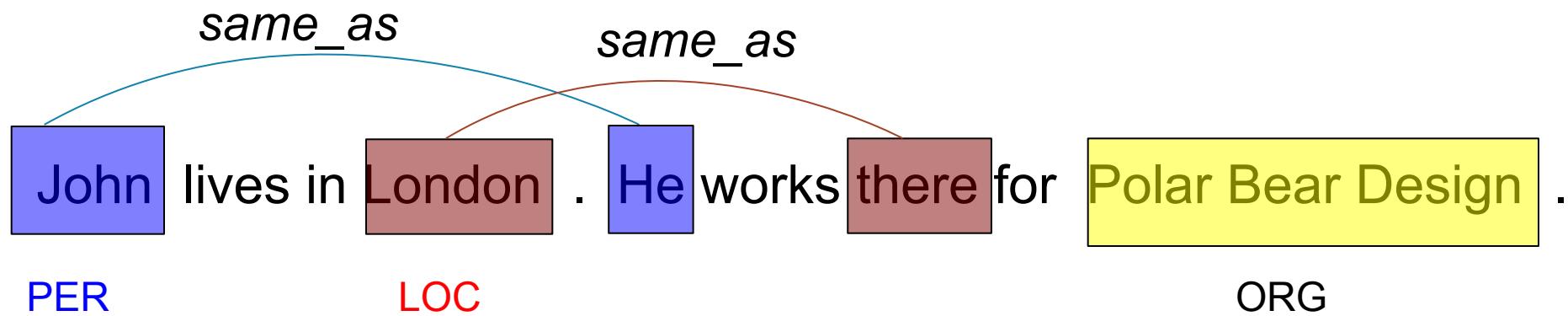
# BASIC NE RECOGNITION

Process gazetteer and grammar rules...

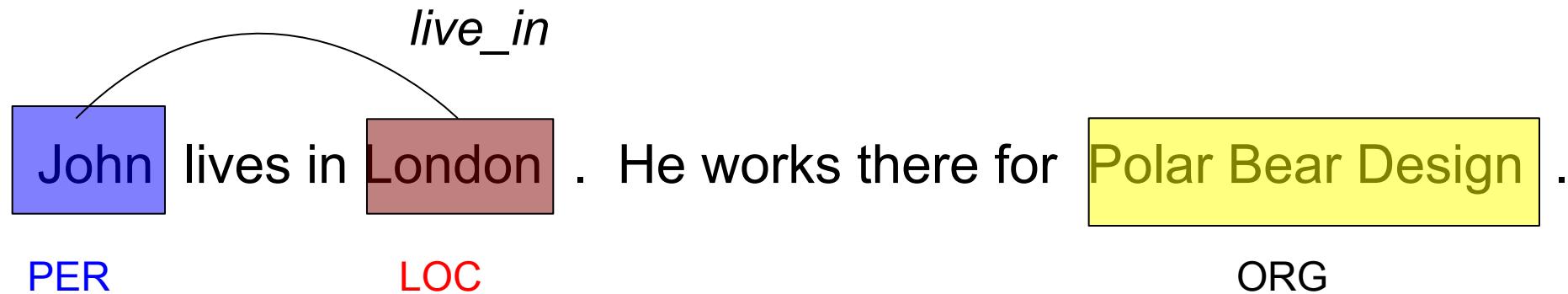
John lives in London . He works there for Polar Bear Design .

PER                    LOC                    ORG

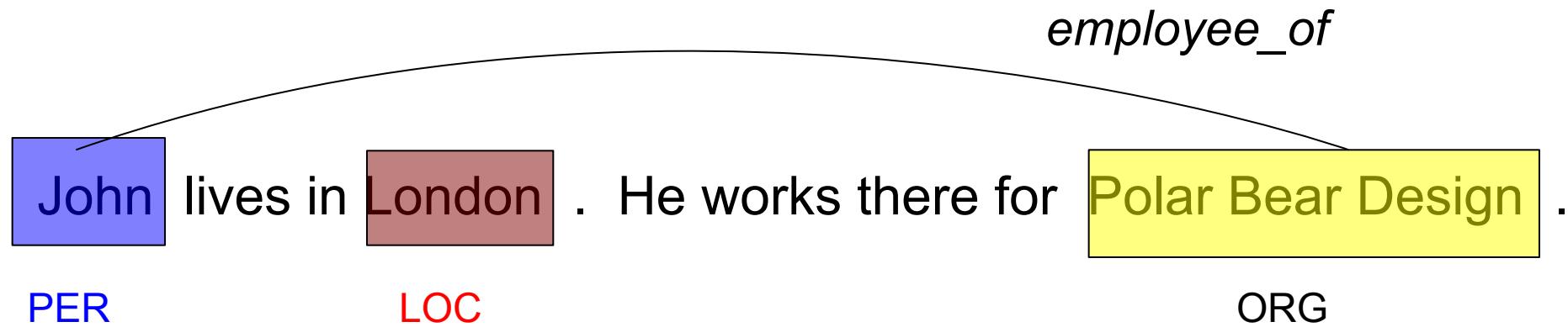
# CO-REFERENCE



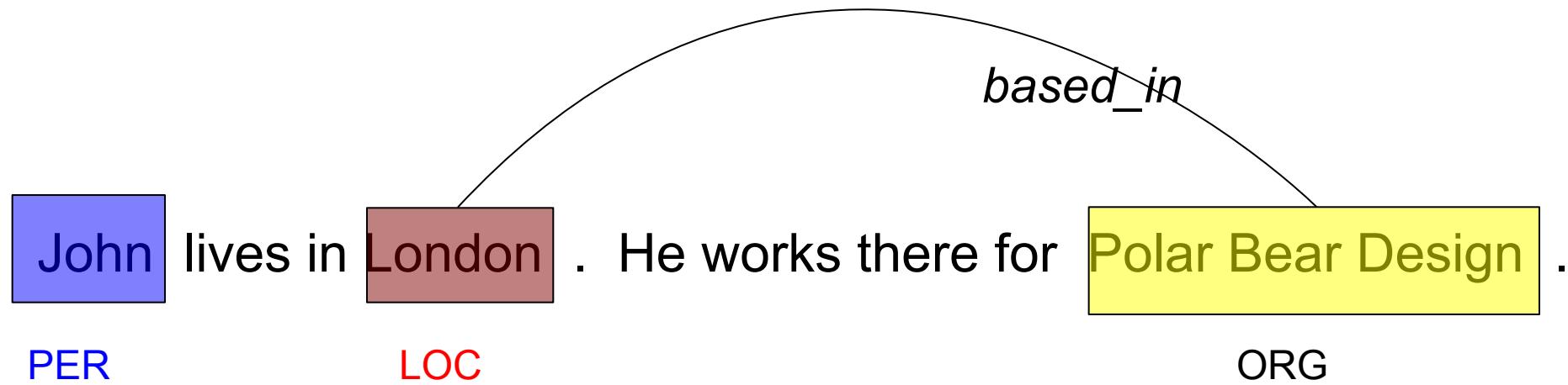
# RELATIONS



# RELATIONS (2)



# RELATIONS (3)





# A CLOSER LOOK AT ANNIE(A NEARLY NEW INFORMATION EXTRACTION SYSTEM)

---



# ABOUT THIS PART

- This part of the tutorial will look have a closer look at ANNIE and its components.
- We will create our own ANNIE-like application.
- As before, this tutorial will be a hands-on session with some explanation as you go.
  - We will use a corpus of news texts in the file [annie-hands-on.zip](#)<sup>1</sup>. Unzip this file if it isn't already.
  - Things for you to try yourself are in red.

1 <http://gate.ac.uk/sale/talks/gate-course-jun19/module-1-ie-and-eval/annie-hands-on.zip>

# WHAT'S IN ANNIE?

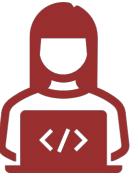
**ANNIE is a ready made collection of PRs that extract information from unstructured text.**

ANNIE contains a set of core PRs which forms a pipeline which consists of:

- Tokeniser
- Sentence Splitter
- POS tagger
- Gazetteers
- Named entity tagger (JAPE transducer)
- Orthomatcher (orthographic coreference)

There are also other useful PRs, which are not used in the default application, but can be added if necessary, e.g.

- NP chunker (in the Tagger:NP Chunking plugin)
- ANNIE VP Chunker (in the Tools plugin)



# LOADING AND RUNNING ANNIE

Because ANNIE is a ready-made application, we can just load it directly from the menu.

- Click the  icon from the top GATE menu.

OR

*File → Ready Made Applications → ANNIE → ANNIE*

OR

right-click **Applications** → *Ready Made Applications → ANNIE → ANNIE*

- Create a new corpus and Populate it from the “news-texts” director.
- Run ANNIE and inspect the annotations.
- You should see a mixture of Named Entity annotations (Person, Location etc.) and some other linguistic annotations (Token, Sentence etc.).

# LET'S LOOK AT THE PRs

Each PR in the ANNIE pipeline **creates some new annotations or modifies existing ones.**

---

**Document Reset** → removes annotations.

---

**Tokeniser** → **Token** annotations.

---

**Gazetteer** → **Lookup** annotations.

---

**Sentence Splitter** → **Sentence, Split** annotations.

---

**POS tagger** → adds **category** features to **Token** annotations.

---

**NE transducer** → **Date, Person, Location, Organisation, Money, Percent** annotations.

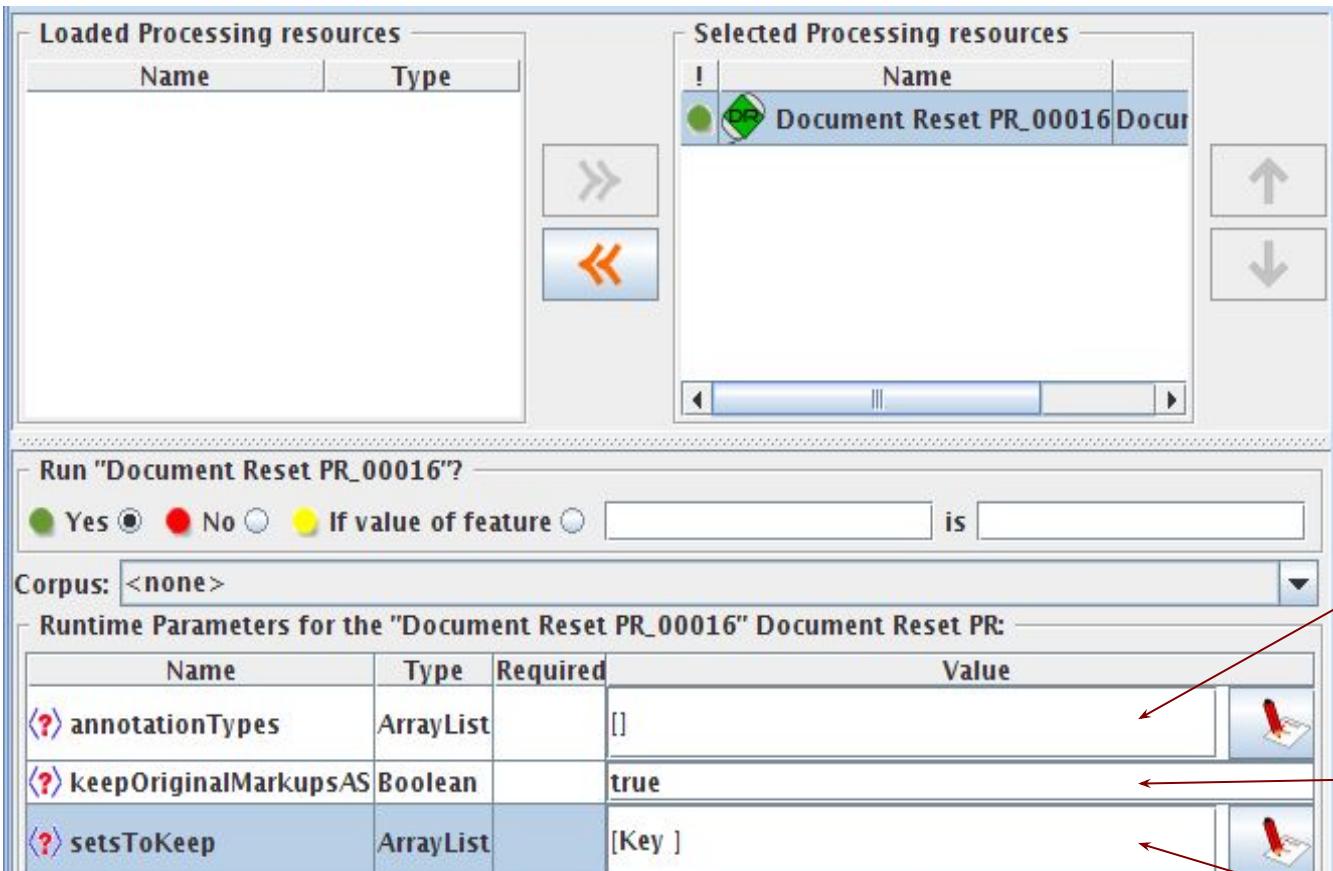
---

**Orthomatcher** → adds **match** features to NE annotations.

# DOCUMENT RESET

- This PR **should go at the beginning** of (almost) every application you create.
- It **removes annotations created previously**, to prevent duplication if you run an application more than once.
- It **does not remove the Original markup set**, by default.
- By default it also **keeps the “Key” set** (by convention the set used for evaluation).
- You can configure it to keep any other annotation sets you want, or to remove particular annotation types only.

# DOCUMENT RESET PARAMETERS



Specify any specific annotations to remove. By default, remove all.

Keep Original Markups set

Keep Key set

# TOKENISER

Splits the text into very simple tokens such as numbers, punctuation and words of different types.

- Produces **Token** and **SpaceToken** annotations with **kind**, **orthography**, **length** and **string** features.
- **kind** can be:
  - word, number, symbol or punctuation.
- **orth (orthography)** can be:
  - upperInitial - initial letter is uppercase, rest are lowercase
  - allCaps - all uppercase letters
  - lowerCase - all lowercase letters
  - mixedCaps - any mixture of upper and lowercase letters not in the above categories
- **length**: number of chars in the token
- **string**: text of the token

# DOCUMENT WITH TOKENS

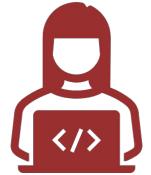
The screenshot shows a user interface for a document annotation tool. At the top, there is a navigation bar with tabs: Annotation Sets, Annotations List, Annotations Stack, Class, Co-reference Editor, Instance, Text, and a search icon. Below the navigation bar, the main area displays a document with several words highlighted in green boxes, indicating they have been annotated as tokens. The document text includes:  
Union Appeals For Talks To End BA Strike  
Skip to navigation | Skip to content |  
Home | Contact Us | News Search;  
HubPage  
Airwise News  
Airport Guide  
Airwise Travel  
Search  
Union Appeals For Talks To End BA Strike  
March 22, 2010  
Union leaders on Sunday called for talks with British Airways bosses to end strike action by cabin crew that has led to the cancellation of hundreds of flights and disrupted travel plans for thousands of passengers.

On the right side of the interface, there is a sidebar titled "Annotations" which lists various annotation types with corresponding color-coded squares:  
Date (purple)  
FirstPerson (red)  
JobTitle (purple)  
Location (blue)  
Lookup (yellow)  
Money (red)  
Organization (red)  
Percent (green)  
Person (green)  
Sentence (purple)  
SpaceToken (pink)  
Split (teal)  
Title (green)  
Token (green, checked)  
Unknown (blue)

At the bottom left, there is a table titled "Annotations" with columns "Type" and "Features". The table lists five tokens with their respective details:

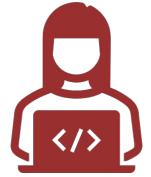
Type	Features
Token	{category=NNP, kind=word, length=5, orth=upperInitial, string=Union}
Token	{category>NNPS, kind=word, length=7, orth=upperInitial, string=Appeals}
Token	{category=IN, kind=word, length=3, orth=upperInitial, string=For}
Token	{category=NNS, kind=word, length=5, orth=upperInitial, string=Talks}
Token	{category=TO, kind=word, length=2, orth=upperInitial, string=To}

# LOOKING AT TOKENS



1. **Tidy up GATE by removing all resources** and applications (or just restart GATE)
2. **Load the news text** hands-on corpus
3. **Create a new application** (corpus pipeline)
4. Load a **Document Reset** and an **ANNIE English Tokeniser**
5. **Add them (in that order)** to the application and run on the corpus
6. **View the Token and SpaceToken annotations**
7. What different values of the “kind” feature do you see?

# SENTENCE SPLITTER



- The default splitter finds sentences based on Tokens.
- Creates **Sentence** annotations and **Split** annotations on the sentence delimiters.
- Uses a gazetteer of abbreviations etc. and a set of JAPE grammars (you will learn in module 2) which find sentence delimiters and then annotate sentences and splits.
- Load an **ANNIE Sentence Splitter PR** and add it to your application **(at the end)**
- **Run the application** and view the results

# DOCUMENT WITH SENTENCES

The screenshot shows the Stanford NLP Annotation Tool interface. The top navigation bar includes tabs for Annotation Sets, Annotations List, Annotations Stack, Class, Co-reference Editor, Instance, Text, and a search icon. The main workspace displays five paragraphs of text, each highlighted with a purple background, indicating selected annotations. To the right is a sidebar containing a list of annotation types with corresponding color-coded squares: Date (light blue), FirstPerson (red), JobTitle (purple), Location (blue), Lookup (yellow), Money (orange), Organization (pink), Percent (green), Person (light green), Sentence (purple, checked), SpaceToken (pink), Split (yellow), Title (light green), Token (light green), Unknown (light blue), and Original markups (grey). Below the sidebar is a table titled 'Annotations Stack' with columns for Type and Features, listing five Sentence annotations, each with an empty feature list ({}).

Type	Features
Sentence { }	

# PART-OF-SPEECH (POS) TAGGER

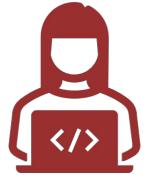
- ANNIE POS tagger is a Java implementation of Brill's transformation based tagger
- Trained on The Wall Street Journal, uses Penn Treebank tagset\*.
- Adds **category** feature to **Token** annotations
- Requires **Tokeniser** and **Sentence Splitter** to be run first

\* [https://www.ling.upenn.edu/courses/Fall\\_2003/ling001/penn\\_treebank\\_pos.html](https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html)

JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP\$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	<i>to</i>
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle

# MORPHOLOGICAL ANALYSER

- Not an integral part of ANNIE, but **can be found in the Tools plugin** as an “added extra”
- Generates **root** feature on **Token** annotations
- **Requires Tokeniser** to be run first
- Requires **POS tagger** to be run first if the **considerPOSTag** parameter is set to true



- Add an **ANNIE POS Tagger** to your app.
- Add a **GATE Morphological Analyser** after the POS Tagger  
(If this PR is not available, load the Tools plugin first).
- **Re-run** your application.
- **Examine the features** of the **Token** annotations.
- New features of **category** and **root** have been added.

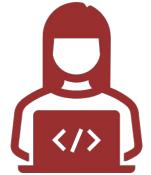
# Gazetteers

# GAZETTEERS

**Gazetteers are plain text files containing lists of names** (e.g. cities, rivers, people, ...). These lists are used to find occurrences of these names in text.

- The ANNIE gazetteer has about 60,000 entries arranged in 80 lists.
- Each list reflects a certain category, e.g. airports, cities, first names etc.
- List entries might be entities or parts of entities, or they may contain contextual information (e.g. job titles often indicate people).
- Each gazetteer has an index file listing all the lists, plus features of each list (**majorType**, **minorType**, and **language**).
- Gazetteers generate **Lookup** annotations with relevant features corresponding to the list matched.
- Lookup annotations are used primarily by the NE transducer.
- Lists can be modified either internally using the Gazetteer Editor, or externally in your favourite editor.

# RUNNING THE ANNIE GAZETTEER

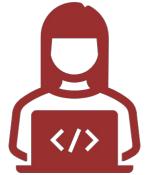


Different kinds of gazetteer are available.

Here, we'll look at the **default ANNIE gazetteer**.

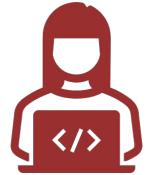
- **Add the ANNIE Gazetteer PR to the end of your pipeline**
- **Re-run the pipeline**
- Look for “**Lookup**” annotations and examine their features

# ANNIE GAZETTEER - CONTENTS



- **Double click on the ANNIE Gazetteer PR** (under Processing Resources in the left hand pane) to open it.
- **Make sure “Gazetteer Editor” is selected** from the bottom tab.
- In the left hand pane (linear definition) you see the index file containing **all the lists**.
- In the right hand pane you see **the contents of the list** selected in the left hand pane.
- **The entries are read-only.**
- To edit ANNIE resources, we first need to **make a copy of them** that we can change.

# EDITING ANNIE GAZETTEER CONTENTS

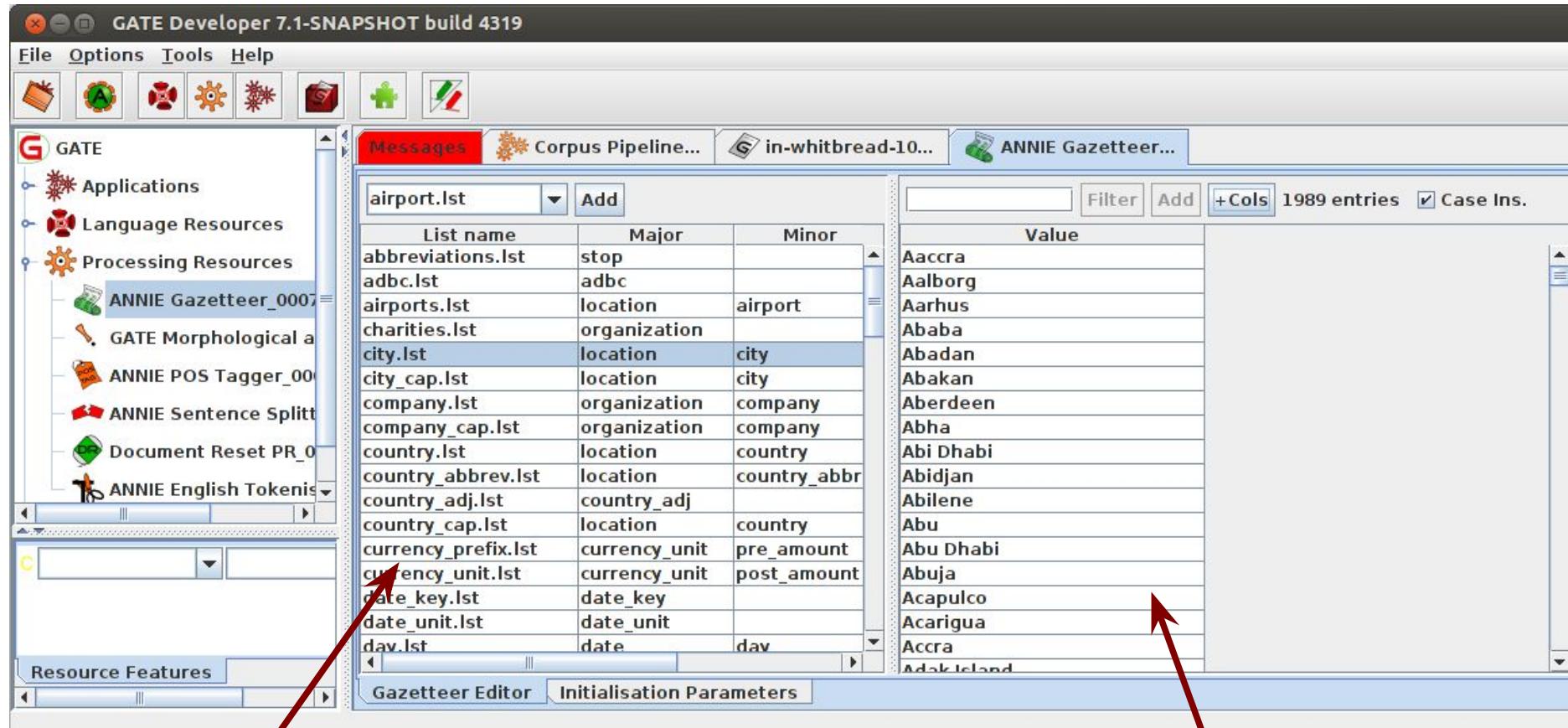


1. **Close ANNIE Gazetteer from PRs.**
2. If you didn't download ANNIE before, go to the plugin manager and **select Annie**, **click the download button**  and **save it to a local folder**.
3. **Right-click PR and select ANNIE Gazetteer.**
4. Click the ***listsURL*** browse button 
5. Select the **file** tab, go to the location you downloaded ANNIE, go to the **gazetteer directory** and select **lists.def** and click OK.
6. Double click **ANNIE Gazetteer** in PR.

Now **each entry can be edited** by clicking in the box and typing.

New entries can be added by typing in the “**New list**” or “**New entry**” box respectively.

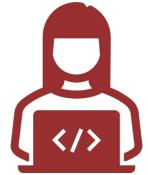
# GAZETTEER EDITOR



definition file  
entries

entries for selected list

# MODIFYING THE DEFINITION FILE



add a new list

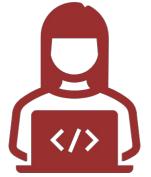
edit an existing list name by typing here

delete a list by right clicking on an entry and selecting 'Delete'

edit the major and minor Types by typing here

List name	Major	Minor
abbreviations.lst	stop	
adbc.lst	adbc	
airports.lst	location	airport
charities.lst	organization	
city.lst	location	city
city_cap.lst	location	city
company.lst	organization	company
company_cap.lst	organization	company
country.lst	location	country
country_abbrev.lst	location	country_abbr
country_adj.lst	country_adj	
country_cap.lst	location	country
currency_prefix.lst	currency_unit	pre_amount
currency_unit.lst	currency_unit	post_amount
date_key.lst	date_key	
date_unit.lst	date_unit	
day.lst	date	day

# MODIFYING A LIST



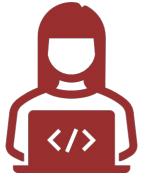
add a new entry  
by typing here

edit an  
existing entry  
by typing here

Value
Aaccra
Aalborg
Aarhus
Ababa
Abadan
Abakan
Aberdeen
Abha
Abi Dhabi
Abidjan
Abilene
Abu
Abu Dhabi
Abuja
Acapulco
Acarigua
Accra
Adak Island

Delete an entry by  
right clicking and  
selecting “Delete”

# EDITING GAZETTEER LISTS



- Click on any list to see the entries.
- Note that some lists are not very complete!
- **Try adding, deleting and editing** existing lists, or the list definition file.
- To save an edited gazetteer, **use *Ctrl-S* shortcut** or right click on the gazetteer name in the tabs at the top or in the resources pane on the right, and select “**Save and Reinitialise**” before running the gazetteer again.
- **Try adding a new word** from a document you have loaded (that is not currently recognised as a Lookup) into the gazetteer, re-run the gazetteer and check the results.

# EDITING GAZETTEERS OUTSIDE GATE

- You can also edit both the definition file and the lists outside GATE, in a text editor
- If you choose this option, **you will need to reinitialise the gazetteer** in GATE before running it again
- **To reinitialise** any PR, **right click on its name** in the Resources pane and **select “Reinitialise”**

# LIST ATTRIBUTES

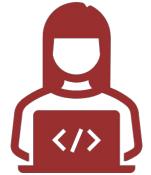
When something in the text matches a gazetteer entry, a **Lookup** annotation is created, with various features and values.

- The ANNIE gazetteer has the following default feature types: **majorType**, **minorType**, **language**.
- For example, the “city” list has a majorType “location” and minorType “city”, while the “country” list has “location” and “country” as its types.
- Later, in the JAPE grammars, we can refer to all Lookups of type location, or we can be more specific and refer just to those of type “city” or type “country”.

# NE TRANSDUCER

- Gazetteers can be used to find terms that suggest entities.
- However, the entries can often be ambiguous.
  - “May Jones” vs “May 2010” vs “May I be excused?”
  - “Mr Parkinson” vs “Parkinson's Disease”.
  - “General Motors” vs. “General Smith”.
- Hand-crafted grammars can be used to **define patterns over the Lookups and other annotations.**
- These patterns can help disambiguate, and they can combine different annotations, e.g. Dates can be comprised of {day} + {number} + {month} Lookup annotations.
- NE transducer consists of a number of grammars **written in the JAPE language.** Module-2 will be **devoted to JAPE.**

# ANNIE NE TRANSDUCER



- **Load an ANNIE NE Transducer PR**
- **Add it to the end of the application**
- **Run the application**
- **Look at the annotations**
- You should see some new annotations such as **Person, Location, Date** etc.
- These will have features showing more specific information (e.g. what kind of location it is) and the rules that were fired (for ease of debugging)

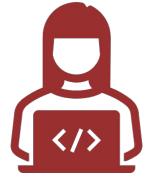
# USING CO-REFERENCE

- Different expressions may refer to the same entity.
- Orthographic co-reference module (orthomatcher) matches proper names and their variants in a document.
- *[Mr Smith]* and *[John Smith]* will be matched as the same person.
- *[International Business Machines Ltd.]* will match *[IBM]*.

# ORTHOMATCHER PR

- Performs co-reference resolution based on orthographical information of entities
- Produces a list of annotation IDs that form a **co-reference “chain”**
- List of such lists stored as a **document feature** named **“MatchesAnnots”**
- Improves results by assigning entity type to previously unclassified names, based on relations with classified entities
- Classification of unknown entities very useful for surnames which match a full name, or abbreviations,  
e.g. “Bonfield” **<Unknown>** will match “Sir Peter Bonfield” **<Person>**

# LOOKING AT CO-REFERENCE



- Add a new PR: **ANNIE OrthoMatcher**.
- Add it **to the end** of the application.
- **Run** the application.
- In a document view, **open the co-reference editor** by clicking the button above the text.

All the documents in the corpus should have some co-reference, but some may have more than others.

# CO-REFERENCE EDITOR

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text 

Seven UK airlines including British Airways, Virgin Atlantic, BMI British Midland and EasyJet, on Friday took over control of the air traffic control system, completing one of the government's most controversial public-private partnership deals.

Completion of the National Air Traffic Services deal comes at a critical time for the government as it tries to push through the PPP for the London Underground.

The sale to a strategic investor of a 46 per cent stake in Nats is the first time in Europe that management control of en route air traffic services has passed into private hands.

It has been carried out despite a pledge by Labour before the 1997 general election that UK air was "not for sale."

Under the terms of the deal, which was approved by the European competition authorities in May, the government has retained a 49 per cent stake and a golden share, while a 5 per cent stake is to be allocated to Nats' 5,700 staff.

The Airline Group, which also includes the charter carriers Airtours International Airways, Britannia Airways and Monarch Airlines, is paying GBP50m (\$71m) to acquire the 46 per cent stake.

Total government proceeds from the deal amount to about GBP800m, with the lion's share of the funds coming from new debt raised by Nats. The Airline Group has agreed financing facilities for Nats with a group of banks led by Barclays and Abbey National.

Completion of the deal has come about two months behind the original schedule announced at the end of March.

It is understood that negotiations were held up by concerns expressed by the banks financing the deal about revised traffic forecasts presented by Nats after the selection of the Airline Group as the government's partner was announced at the end of March.

The Airline Group is taking over Nats at a difficult time with air traffic control capacity under increasing pressure from rising air traffic volumes.

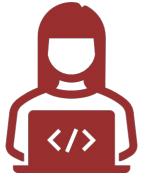
Sets : Default  Types : Organization  Show

Co-reference Data

Default

- National Air Traffic Services
- Airline Group
- UK
- Swanwick
- March

# USING THE CO-REFERENCE EDITOR



- **Select the annotation set** you wish to view (select Default set for now)
- A list of all the co-reference chains that are based on annotations in the currently selected set is displayed
- **Select an item in the list** to highlight all the member annotations of that chain in the text (you can select more than one at once)
- **Hovering over** a highlighted annotation in the text **enables you to delete** an item from the co-reference chain
- **Try it!**

# EVALUATION

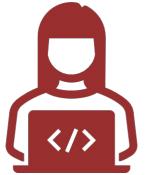


“We didn’t underperform. You over expected.”

# INTRODUCTION TO EVALUATION

- Evaluation of NLP tools is very important because we need to know how well our tools are performing.
  - Is it actually worth developing an automatic tool to perform a task?
  - Especially in GATE, there is often a choice of which tool to use for a job (e.g. multiple parsers) so we might want to know which one is best.
  - We need to know whether changes we make to the tools will improve or harm our system: e.g. making components case-insensitive might improve Recall but harm Precision
  - We will look at what evaluation metrics to use for NLP, and some tools to perform evaluation.

# EVALUATION EXERCISES: PREPARATION



- Restart GATE, or **close all documents** and PRs to tidy up
- **Load the annie-hands-on/news-texts** into a corpus
- **Take a look at the annotations.**
- There is an annotation set called “**Key**”. This is a set of annotations against which we want to evaluate ANNIE. In practice, **they can be created by human annotators, or by another application.**
- **Load ANNIE and run it**
- **You should have annotations in the Default set from ANNIE, and in the Key set, against which we can compare them.**

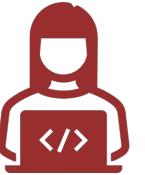
# ANNOTATION DIFF

- Graphical comparison of 2 sets of annotations
- Visual diff representation, like tkdiff
- Compares one document at a time, one annotation type at a time

# Annotations are like squirrels...



Annotation Diff helps with “spot the difference”



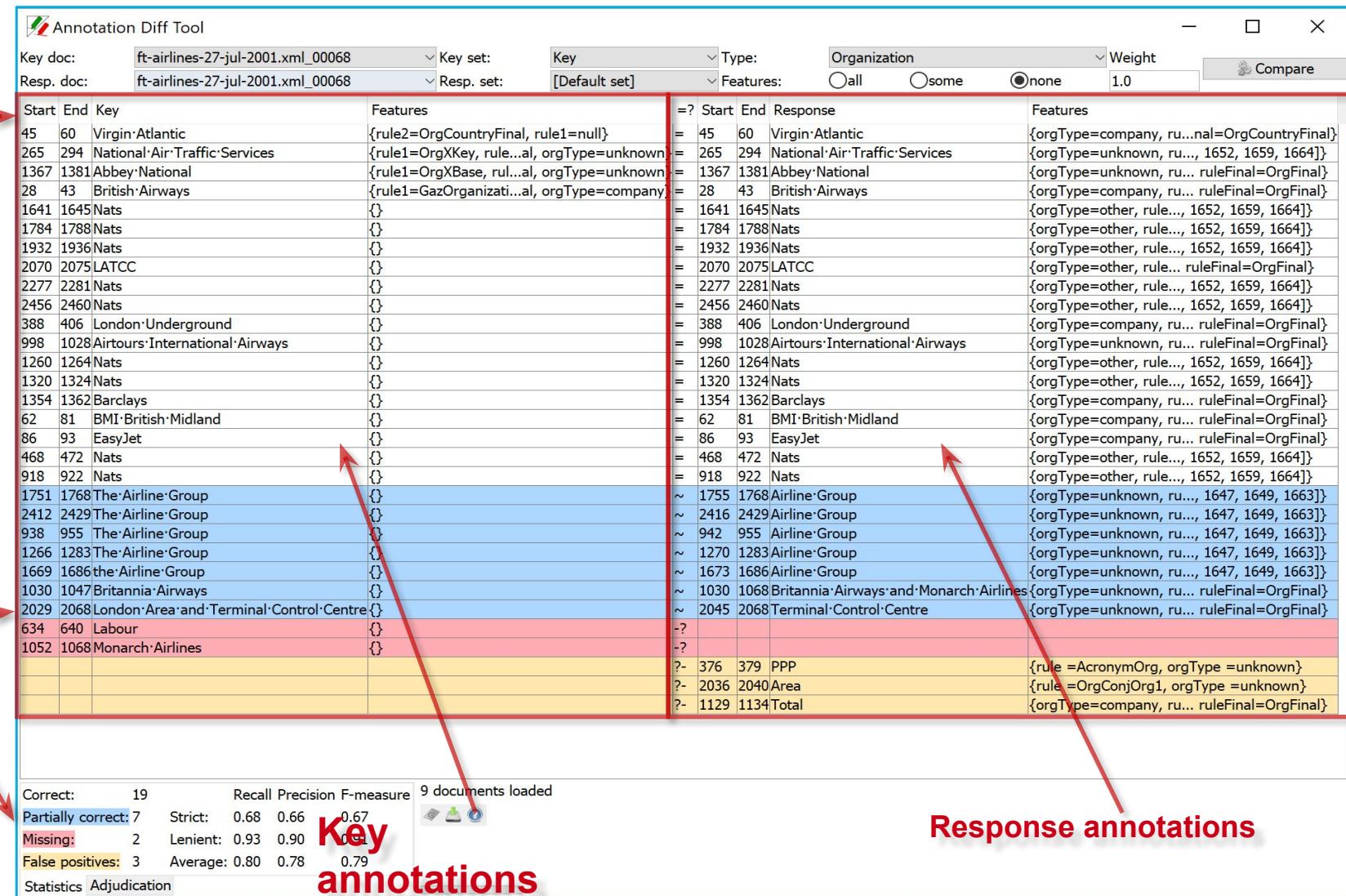
# ANNOTATION DIFF EXERCISE

- Open the document “**ft-airlines-27-jul-2001.xml**”
- Open the **Annotation Diff** (Tools → Annotation Diff or **click icon** - For the **Key** set (may contain the manual annotations) select **Key** annotation set
- For the **Response** set (containing annotations from ANNIE) select the **Default** annotation set
- For the **Type** option select the **Organization** annotation.
- Click on “Compare”
- Scroll down the list, to see correct, partially correct, missing and false positive annotations

# COMPARING THE ANNOTATIONS

You can sort the columns however you like.

colour codes indicate whether the annotation pair shown are correct, partially correct, missing (false negative) or false positive.



# MEASURING SUCCESS

In IE, we classify the annotations produced in one of 4 ways:

**Correct** = things annotated correctly

e.g. annotating “Sheffield” as a Location

**Missing** = things not annotated that should have been

e.g. not annotating “Sheffield” as a Location

**False positive** = things annotated wrongly

e.g. annotating “Sheffield United F.C.” as a Location.

**Partially correct** = the annotation type is correct, but the span is wrong

e.g. annotating just “Trump” as a Person (too short) or annotating  
“Unfortunately Donald Trump” as a Person (too long)

# FINDING PRECISION, RECALL AND F-MEASURE

Annotation Diff Tool

Key doc: ft-airlines-27-jul-200... Key set: Key Type: Organization Weight: Compare

Resp. doc: ft-airlines-27-jul-200... Resp. set: [Default set] Features:  all  some  none 1.0

Start	End	Key	Features	=?	Start	End	Key
1932	1936	Nats	{}	=	1932	1936	Nats
2456	2460	Nats	{}	=	2456	2460	Nats
2070	2075	LATCC	{}	=	2070	2075	LATCC
1354	1362	Barclays	{}	=	1354	1362	Barclays
1784	1788	Nats	{}	=	1784	1788	Nats
1751	1768	The·Airline·Group	{}	~	1755	1768	Airline·Gro
938	955	The·Airline·Group	{}	~	942	955	Airline·Gro
1669	1686	the·Airline·Group	{}	~	1673	1686	Airline·Gro
2412	2429	The·Airline·Group	{}	~	2416	2429	Airline·Gro
1266	1283	The·Airline·Group	{}	~	1270	1283	Airline·Gro
1052	1068	Monarch·Airlines	{}	~	1030	1068	Britannia·A
2029	2068	London·Area·and·Terminal·Control·Centre	{}	~	2045	2068	Terminal·C
634	640	Labour	{}	?			
1030	1047	Britannia·Airways	{}	?			
				?-	2029	2040	London·Are
				?-	2386	2395	Hampshire

Correct: 19      Partially correct: 7      Missing: 2      False positives: 2

Recall Precision F-measure

Strict:	0.68	0.68	0.68
Lenient:	0.93	0.93	0.93
Average:	0.80	0.80	0.80

10 documents loaded

Statistics Adjudication

scores displayed

# Precision

**Correct** = things annotated correctly

**Missing** = things not annotated that should have been

**False positive** = things annotated wrongly

- How many of the entities your application found were correct?

$$Precision = \frac{Correct}{Correct + \text{False positive}}$$

# Recall

**Correct** = things annotated correctly

**Missing** = things not annotated that should have been

**False positive** = things annotated wrongly

- How many of the entities that exist did your application find?
- Sometimes recall is called **coverage**

$$Recall = \frac{Correct}{Correct + Missing}$$

# F-MEASURE



Precision and recall tend to **trade off against one another**.

\* If you specify your rules precisely to improve precision, you may get a lower recall.

\* If you make your rules very general, you get good recall, but low precision.



This makes it difficult to compare applications, or to check whether a change has improved or worsened the results overall.



F-measure combines precision and recall into one measure.

# F-MEASURE

- Also known as the “harmonic mean”.
- Usually, precision and recall are equally weighted.
- This is known as **F1**.
- To use F1, set the value of the F-measure weight to 1, this is the default setting in Annotation Diff tool.

$$F = 2 \cdot \left( \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \right)$$

# ANNOTATION DIFF DEFAULTS TO F1

Annotation Diff Tool

Key doc: ft-airlines-27-jul-200... Key set: Key Type: Organization Weight: Compare

Resp. doc: ft-airlines-27-jul-200... Resp. set: [Default set] Features:  all  some  none 1.0

Start	End	Key	Features	=?	Start	End	Key
1932	1936	Nats	{}	=	1932	1936	Nats
2456	2460	Nats	{}	=	2456	2460	Nats
2070	2075	LATCC	{}	=	2070	2075	LATCC
1354	1362	Barclays	{}	=	1354	1362	Barclays
1784	1788	Nats	{}	=	1784	1788	Nats
1751	1768	The·Airline·Group	{}	-	1755	1768	Airline·Gro
938	955	The·Airline·Group	{}	-	942	955	Airline·Gro
1669	1686	the·Airline·Group	{}	-	1673	1686	Airline·Gro
2412	2429	The·Airline·Group	{}	-	2416	2429	Airline·Gro
1266	1283	The·Airline·Group	{}	-	1270	1283	Airline·Gro
1052	1068	Monarch·Airlines	{}	-	1030	1068	Britannia·A
2029	2068	London·Area·and·Terminal·Control·Centre	{}	-~	2045	2068	Terminal·C
634	640	Labour	{}	-?			
1030	1047	Britannia·Airways	{}	-?			
				?-	2029	2040	London·Are
				?-	2386	2395	Hampshire

Correct: 19      Recall: 0.68      Precision: 0.68      F-measure: 0.68

Partially correct: 7      Strict: 0.68      Lenient: 0.93      Average: 0.80

Missing: 2      F-measure weight set to 1

False positives: 2      F-measure weight set to 1

Statistics   Adjudication

10 documents loaded

# STATISTICS CAN MEAN WHAT YOU WANT THEM TO....

How we want to measure partially correct annotations may differ, depending on our goal.

In GATE, there are 3 different ways to measure them

- The most usual way is to consider them to be “**half right**”.
- **Strict:** Only perfectly matching annotations are counted as correct.
- **Lenient:** Partially matching annotations are counted as correct. This makes your scores look better : ))
- **Average:** Strict and lenient scores are averaged (this is the same as counting a half weight for every partially correct annotation).

# STRICT, LENIENT AND AVERAGE

Annotation Diff Tool

Key doc: ft-airlines-27-jul-200... Key set: Key Type: Organization Weight  
Resp. doc: ft-airlines-27-jul-200... Resp. set: [Default set] Features:  all  some  none 1.0 Compare

Start	End	Key	Features	=?	Start	End	Key
1932	1936	Nats	{}	=	1932	1936	Nats
2456	2460	Nats	{}	=	2456	2460	Nats
2070	2075	LATCC	{}	=	2070	2075	LATCC
1354	1362	Barclays	{}	=	1354	1362	Barclays
1784	1788	Nats	{}	=	1784	1788	Nats
1751	1768	The·Airline·Group	{}	~	1755	1768	Airline·Gro
938	955	The·Airline·Group	{}	~	942	955	Airline·Gro
1669	1686	the·Airline·Group	{}	~	1673	1686	Airline·Gro
2412	2429	The·Airline·Group	{}	~	2416	2429	Airline·Gro
1266	1283	The·Airline·Group	{}	~	1270	1283	Airline·Gro
1052	1068	Monarch·Airlines	{}	~	1030	1068	Britannia·A
2029	2068	London·Area·and·Terminal·Control·Centre	{}	~	2045	2068	Terminal·C
634	640	Labour	{}	-?			
1030	1047	Britannia·Airways	{}	-?			
				?-	2029	2040	London·Are
				?-	2386	2395	Hampshire

Correct: 19      Partially correct: 7      Missing: 2      False positives: 2      Recall: 0.68      Precision: 0.68      F-measure: 0.68      Lenient: 0.93      Average: 0.80

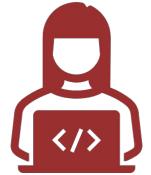
10 documents loaded

Statistics Adjudication

# CORPUS QUALITY ASSURANCE

- **Corpus Quality Assurance tool extends the Annotation Diff functionality to the entire corpus, rather than on a single document at a time.**
- **It produces statistics both for the corpus as a whole (Corpus statistics tab) and for each document separately (Document statistics tab).**
- **It compares two annotation sets, but makes no assumptions about which (if either) set is the gold standard.** It just labels them A and B.
- This is because it can be used to measure **Inter Annotator Agreement (IAA)** where there is no concept of “correct” set.

# TRY OUT CORPUS QUALITY ASSURANCE

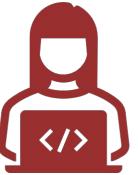


The screenshot shows the GATE Corpus Pipeline interface. On the left, there's a sidebar with various processing resources listed under 'Processing Resources': Batch Learning PR\_0009D, Jape Transducer\_00094, ANNIE OrthoMatcher, NE ANNIE NE Transducer, ANNIE POS Tagger, ANNIE Sentence Splitter, ANNIE Gazetteer, ANNIE English Tokeniser, and Document Reset PR. The 'GATE Corpus\_0001A' resource is selected, highlighted with a yellow background. The main pane displays a list of 14 documents under the heading 'All the documents loaded in the system are in this corpus.' The list includes:

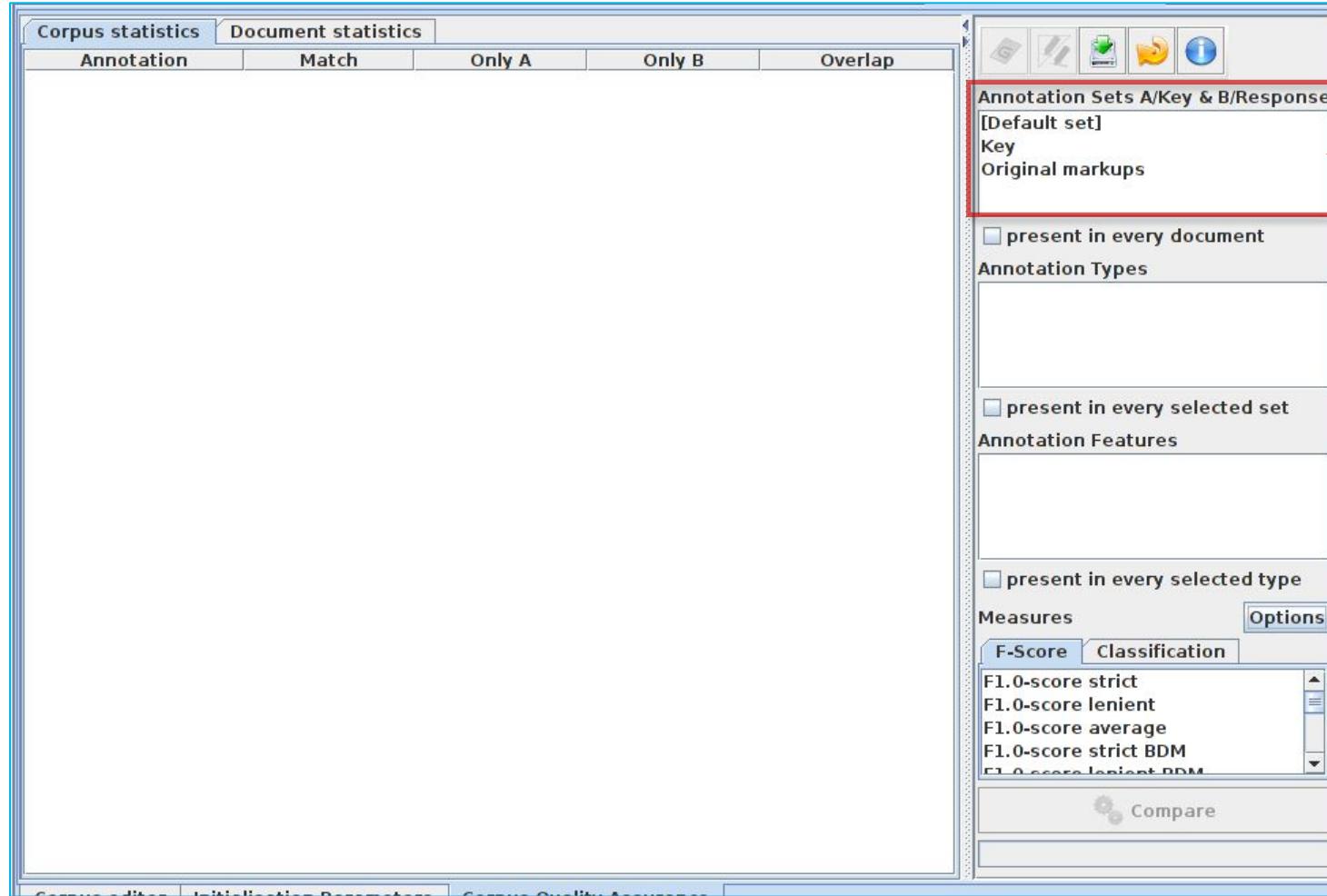
Index	Document name
0	ft-BT-07-aug-2001.xml_0001B
1	ft-BT-briefing-02-aug-2001.xml_0001C
2	ft-BT-loop-01-aug-2001.xml_0001D
3	ft-GKN-09-aug-2001.xml_0001E
4	ft-SSL-10-aug-2001.xml_0001F
5	ft-WestLB-BT-05-aug-2001.xml_00020
6	ft-airlines-27-jul-2001.xml_00021
7	ft-airtours-08-aug-2001.xml_00022
8	ft-bank-of-england-02-aug-2001.xml_00023
9	ft-bank-of-uk-08-Aug-2001.xml_00024
10	ft-bmi-09-may-2001.xml_00025
11	ft-bmi-25-feb-2001.xml_00026
12	ft-bmi-airline-07-aug-2001.xml_00027
13	ft-bt-03-aug-2001.xml_00028
14	ft-bt-26-jul-2001.xml_00029

At the bottom of the interface, there are three tabs: 'Corpus editor', 'Initialisation Parameters', and 'Corpus Quality Assurance'. The 'Corpus Quality Assurance' tab is highlighted with a red border and has a red arrow pointing to it from the right side of the image.

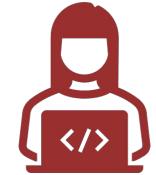
Open your hands-on corpus and click the **Corpus Quality Assurance** tab at the bottom of the Display pane.



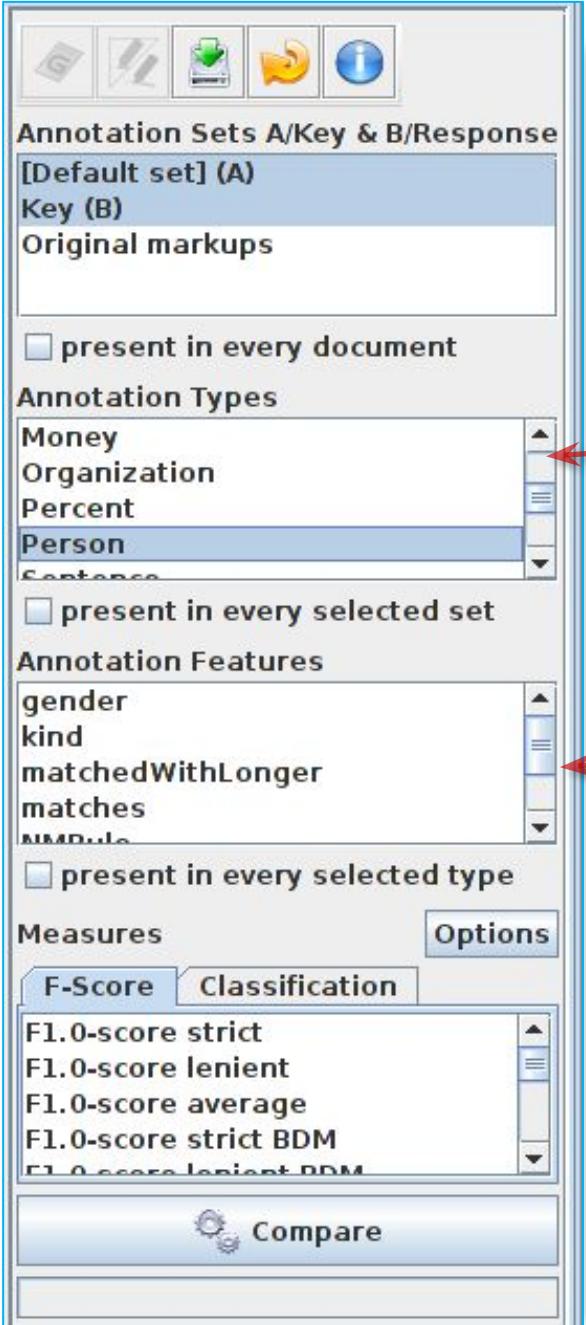
# SELECT ANNOTATION SETS



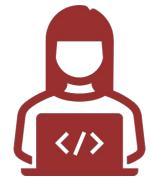
- Select the annotation sets you wish to compare.
- Click on the Key annotation set – this will label it set A.
- Now click on the default annotation set - this will label it set B.



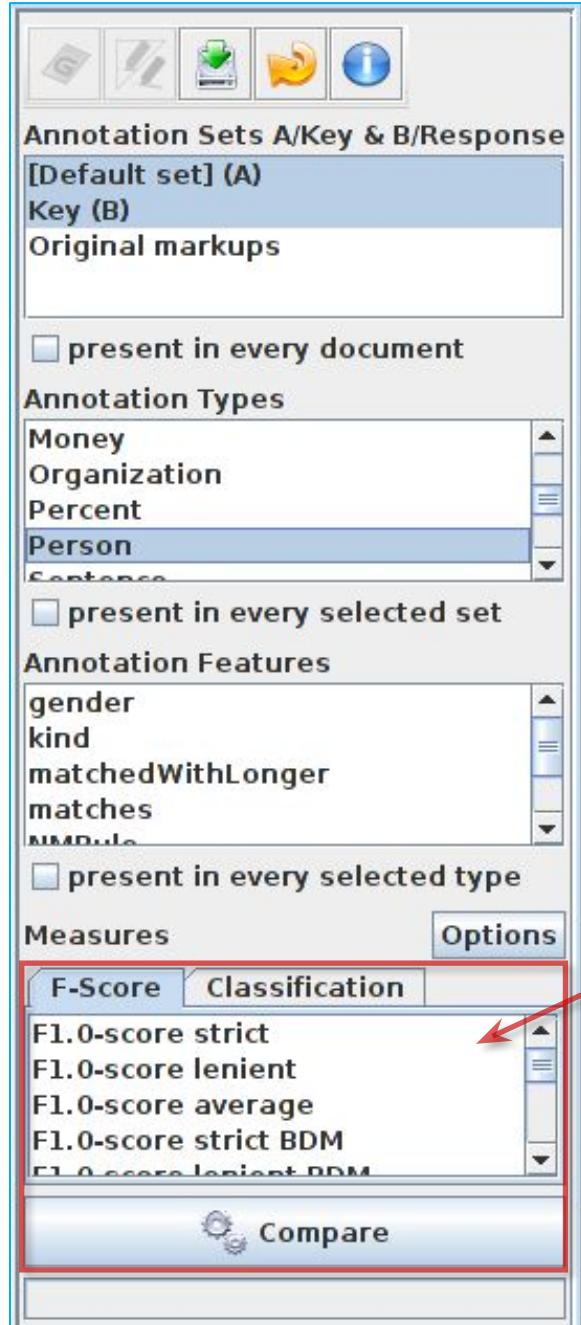
# Select Type



- Select the annotation type to compare (suggestion: select Organisation, Person and Location for now)
- Select the features to include (if any – leave unselected for now)
- You can select as many types and features as you want.



# Select measure



- In the “**Measures**” box, **select the kind of F score** you want “Strict, Lenient, Average” or any combination of them.
- Suggestion: try just “**lenient**” at first.
- **Select Compare**

# Corpus Statistics Tab

Corpus statistics		Document statistics					
Annotation	Match	Only A	Only B	Overlap	Prec.B/A	Rec.B/A	F1.0-l.
Location	55	6	0	5	1.0000	0.9091	0.9524
Organization	76	5	15	11	0.8529	0.9457	0.8969
Person	28	2	1	0	0.9655	0.9333	0.9492
Macro summary					0.9395	0.9294	0.9328
Micro summary	159	13	16	16	0.9162	0.9309	0.9235

- Each annotation type is listed separately
- Precision, recall and F measure are given for each
- Two summary rows provide micro and macro averages

# Micro and Macro Averaging

- Micro averaging treats the entire corpus as one big document, for the purposes of calculating precision, recall and F.
- Macro averaging takes the average of the rows.  
Here: the average **over different annotation types**

# Document Statistics Tab

Corpus statistics	Document statistics							
Document		Match	Only A	Only B	Overlap	Prec.B/A	Rec.B/A	F1.0-l.
ft-airlines-27-jul-2001.xml_00030		28	4	2	7	0.9459	0.8974	0.9211
ft-airtours-08-aug-2001.xml_00031		19	0	0	0	1.0000	1.0000	1.0000
ft-bank-of-england-02-aug-2001.xml_00032		22	2	2	1	0.9200	0.9200	0.9200
ft-bmi-09-may-2001.xml_00033		24	1	2	2	0.9286	0.9630	0.9455
ft-claims-direct-10-aug-2001.xml_00034		21	2	1	0	0.9545	0.9130	0.9333
ft-commerzbank-10-aug-2001.xml_00035		10	2	4	2	0.7500	0.8571	0.8000
ft-equitable-07-auf-2001.xml_00036		9	1	4	0	0.6923	0.9000	0.7826
ft-house-price-08-aug-2001.xml_00037		9	1	0	1	1.0000	0.9091	0.9524
ft-industrial-gloom-07-Aug-2001.xml_00038		17	0	1	3	0.9524	1.0000	0.9756
Macro summary						0.9049	0.9288	0.9145
Micro summary		159	13	16	16	0.9162	0.9309	0.9235

- Each document is listed separately.
- Precision, recall and F measure are given for each.
- Two summary rows provide micro and macro (here: **over documents**) averages.

# SUMMARY



This session has been devoted to IE and ANNIE



You should now have  
a basic  
understanding of:  
what IE is and

how to load and run ANNIE,  
what each of the ANNIE components do,  
how to modify ANNIE components,



Evaluation using Annotation Diff and Corpus QA.

# FUN EXTRA TASK



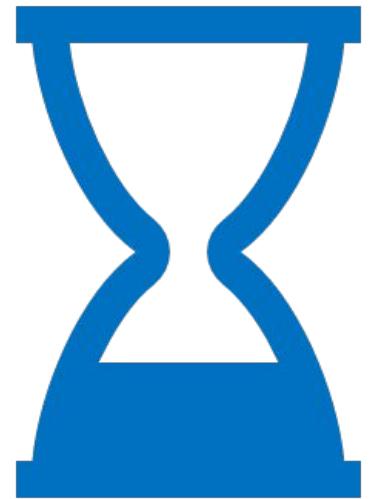
- If you have time, you can try annotating a document yourself with named entities and then comparing how you did with the existing Key annotation set
- Reminder: to annotate a document, make sure the right annotation set is selected with the mouse (we suggest adding a new one with your name) and then highlight the text you want to annotate. A popup window will appear, letting you select the annotation type.
- Use one of the evaluation tools to compare how you did!

# THANK YOU!

Any  
Questions?

## OPTIONAL MATERIAL

---



# SAVING DOCUMENTS



Using datastores



Saving documents for  
use outside GATE

# TYPES OF DATASTORES

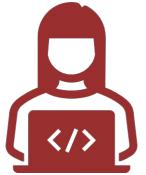
There are 2 types of datastore:

**Serial datastores store** data directly in a directory.

**Lucene datastores** provide a searchable repository with Lucene-based indexing.

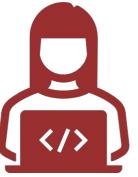
For now, we'll look at serial datastores.

# CREATE A NEW SERIAL DATASTORE



- **Right click “Datastores” from the Resources pane and select “Create Datastore”**
- **Select “Serial Datastore”**
- Create a new empty directory by clicking the “**Create New Folder**” icon and give your new directory a name
- Note: if this icon does not appear, try selecting the Metal Look & Feel (especially Mac users)
- **Select this directory and click “Open”**

Now your datastore is ready to store your documents



# SAVE DOCUMENTS TO THE DATASTORE

- **Right click on your corpus and select “Save to Datastore”**
- **Select the datastore that you just created**
- **Now close the corpus and document**
- **Double click on the name of the datastore in the Resources pane**  
You should see the corpus and document
- **Double click on them to load them back into GATE and view them**

They should contain the annotations you created previously

You can remove things from the datastore **by right clicking** on their name in the datastore and **selecting “Delete”**

You can add several corpora to the same datastore

Note: in general, it's best to save the empty corpus to the datastore and **then** populate it, to avoid keeping a lot of documents in memory

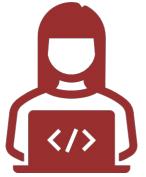
# SAVING DOCUMENTS OUTSIDE GATE

**Datastores can only be used inside GATE**, because they use a GATE-specific format

If you want to **use your documents outside GATE**, you can save them in 2 ways:

- **Gate XML** is a **standoff markup**, a special GATE representation
- **Inline XML** has **inline annotations** (preserving the original format)

# SAVING AS XML

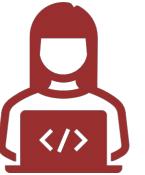


- Load any document from the hands-on material into GATE, then right click on it in the Resources pane
- Select “Save as Gate XML” and select a filename.

In this format, all annotations are appended to the end of the document and the location for each annotation is marked by a tag in the body of the document

Each annotation has a unique ID

If you’re curious, load the document into your favourite text editor and have a look at it!



## IF YOU HAVE LOTS OF DOCUMENTS IN A CORPUS...

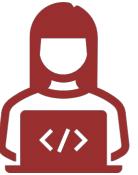
A datastore is the best way to store them, because it uses less memory in GATE when processing

- Delete all corpora and documents in your datastore
- Load a new corpus (*Language Resources* → *New* → *GATE Corpus*)
- Create a new datastore and **save the (empty) corpus** to the datastore
- Now populate your corpus (**right click on corpus** → *Populate*)

You should see the documents appear in your datastore

Your documents will be loaded into the datastore and saved automatically.

- Close and reopen your datastore to check they really were saved!



# SAVE AS INLINE XML

This option will save the document with all the original annotations from HTML or XML documents, and any new annotations that you currently have selected in the document editor.

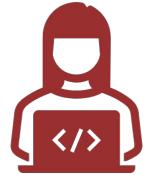
This can be useful for saving only a subset of the annotation types.

Annotations are saved using standard XML tags, with the annotation type as the tag name.

Partially overlapping annotations can not be saved.

- Right click on a document and select “**Inline XML**”
- Enter the annotationSetName
- Select annotationTypes
- Enter a name for the rootElement, e.g. doc
- Make sure the target path in Save To is correct (this can be an issue in windows machines)
- Click OK to save it.

# FURTHER EXERCISES



- Load an HTML or XML document with the `markupAware` parameter set to false and see the difference
- Investigate the `AnnotationStack`
- Play with Advanced Options
- Run an application over documents in a datastore