Spanish Stemmers for Solr

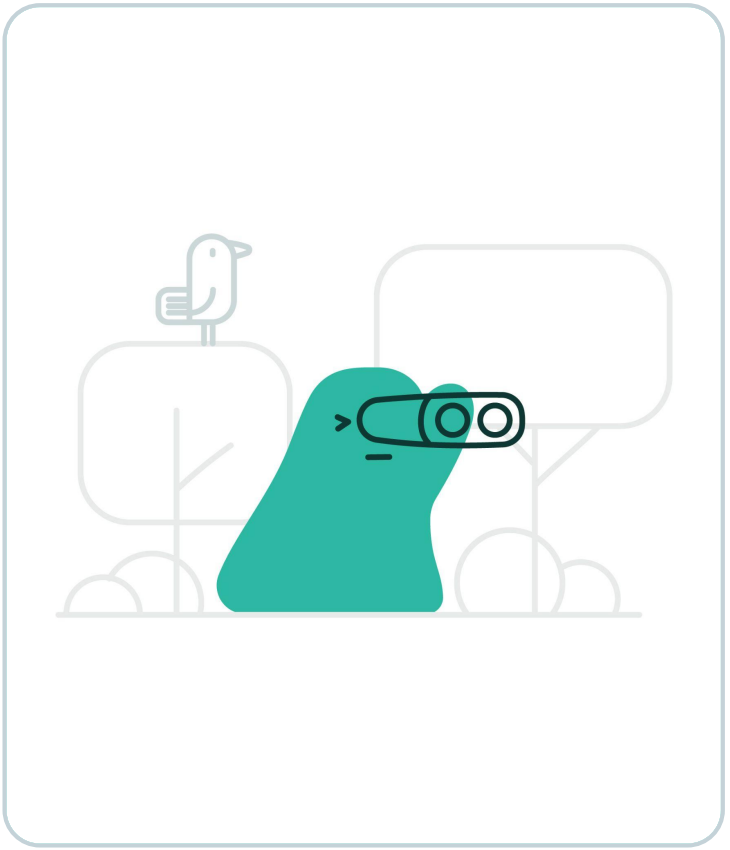
Xavier Sánchez Loro

29/6/2021

wallapop

Outline

1. Available implementations in Solr
2. Spanish Plural Stemmer
3. Stemmers in action
4. Some general guidelines



Current Solr implementations

Spanish Stemmer

- Snowball based stemmer (https://snowballstem.org/)
- General purpose stemmer: stems plural/singular forms, masculine and feminine inflections, verbal forms, adverbs…
- It will increase recall but precision can be drastically reduced
- Lots of collisions between wods with really different meaning
- Use it when recall is more important than precision
- Does not stem plural words of foreign origin
  - i.e. complots, bits, punks, robots

Spanish Light Stemmer

- Algorithmic approach based on the algorithm described in "**Report on CLEF-2001 Experiments**" by Jacques Savoy
- Designed to stem plural to singular form and feminine and masculine inflections to the same root
- It will increase recall but precision can be reduced depending on the use case/information need
- Use it when distinction between singular and plural is not relevant and gender is also not relevant
- Does not stem plural words of foreign origin
  - i.e. complots, bits, punks, robots
- Some collisions between words with really different meaning
  - caro (expensive), cara (face/expensive)
  - barra (bar), barro (mud)

What if we just want to collapse singular and plural whilst still distinguishing between masculine and feminine forms?

We need a new stemmer onl for plurals ;-)

Spanish Plural Stemmer

Custom implementation of a spanish plural stemmer (I)

- Algorithmic approach spanish rules for building plural forms
  - based on rules defined in http://www.wikilengua.org/index.php/Plural_(formaci%C3%B3n)
- Designed to stem just plural to singular form
- Distinguish between masculine and feminine forms
- It will increase recall but precision can be reduced depending on the use case/information need
- Stems plural words of foreign origin
  - i.e. complots, bits, punks, robots
- Support for invariant words: same plural and singular form or plural does not make sense
  - crisis, jueves, lapsus, abrebotellas, etc

Custom implementation of a spanish plural stemmer (II)

- Support for special cases
  - yoes, clubes, itemes, faralaes
- Use it when distinction between singular and plural is not relevant but gender is relevant
- Produces meaningful tokens in form of singular
  - Not strange stems like “amig”
- Preparing it to be released to the community

Stemmers in Action

Word	SpanishStemmer	SpanishLightStemmer	SpanishPluralStemmer	Comments
ases	ases	ases	as	Short plural word
esprais	esprais	esprais	espray	Plural, singular changes to 'y'
cómics	comics	comics	comic	Foreing origin plural
camisas	camis	camis	camisa	Plural, feminine, no masculine form
paces	pac	paz	paz	Plural, singular changes to 'z'
bits	bits	bits	bit	Plural, singular changes to 'y'
cantar	cant	cantar	cantar	Verb

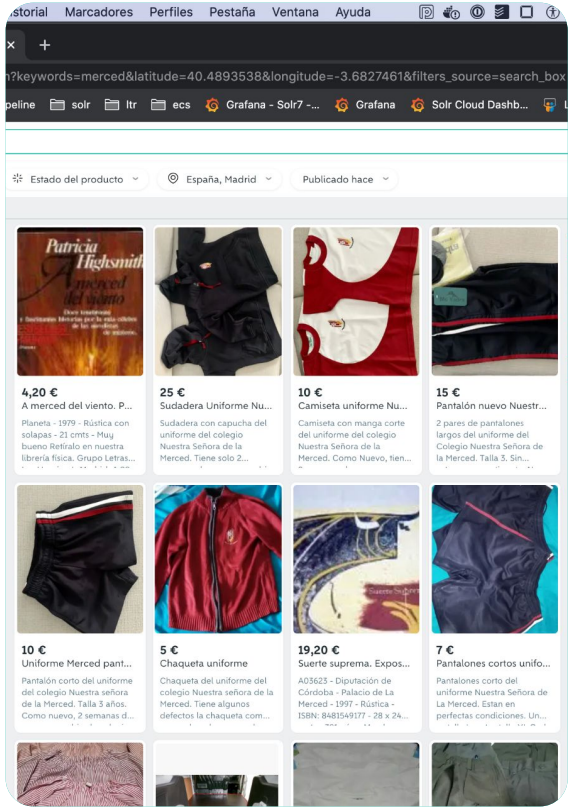
Word	SpanishStemmer	SpanishLightStemmer	SpanishPluralStemmer	Comments
caro/cara	car	car	caro/cara	Singular form (m/f)
abrebotellas	abrebotell	abrebotell	abrebotellas	Invariant
caries	cari	cari	caries	Invariant
amigos/amigas	amig	amig	amigo/amiga	Plural form (m/f)
jueves	juev	juev	jueves	Invariant (name)
osos/osas	osos/osas	osos/osas	oso/osa	Plural form (m/f)

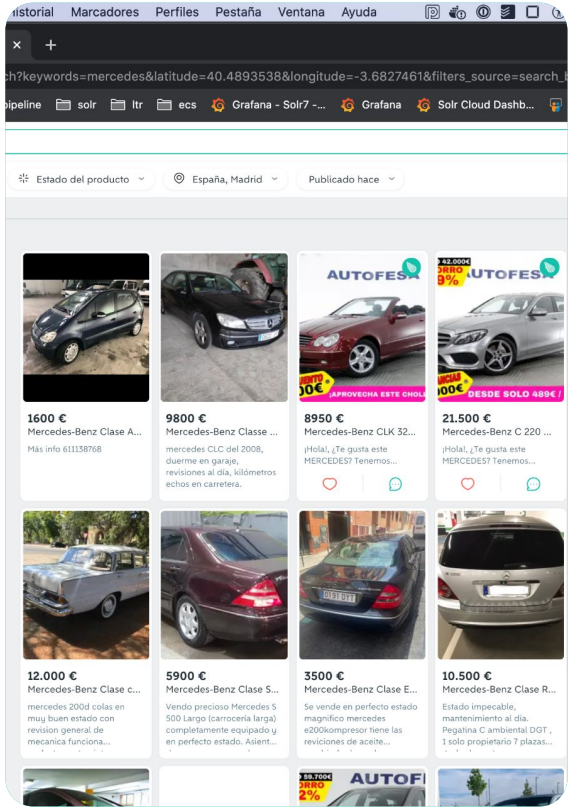
Some General Guidelines

Look for unwanted collisions

Use SetKeywordMarkerFilter for marking tokens as keywords to avoid collisions

- brands, models, names, etc





Choose it for specific purposes

- removing plurals
- removing gender
- increase recall at cost of precision: when you want to see everything Vs I want a specific document or type of document

Objective	Spanish Stemmer	Spanish Light Stemmer	Spanish Plural Stemmer
removing plurals	yes	yes	yes
removing gender	yes	yes	no
stem verbs and related concepts	yes	no but with some collisions	no but with some collisions

Thank you!

Día/mes/año

wallapop