

### 01 INTRODUCTION

Every year the **University of Lleida**, in particular the **Quality and Teaching Planning** unit conducts a lot of satisfaction **surveys** addressed to students, professors, staff, etc.

These surveys include:

Closed questions with valuation form 1 to 5

Free text comments

**Problem:** some of the comments include **offensive expressions**, descriptions of conflicts, errors in professors' scheduling, etc.

I developed an **application** to process these comments.



#### 01 OBJECTIVES



The main objectives of the application are:

- ✓ Importation of surveys and comments
- ✓ Use of Artificial Intelligence (AI) and Natural Language Processing (NLP) to analyze the comments.
- ✓ Visualization of results and manual edition of comments.

#### Other objectives:

- ❖ Use an Agile methodology: **Behaviour Driven Development** (BDD)
- ❖ Test a framework of web development: **Django** and Python
- Explore Natural Language Processing: Spacy library

## 01 QUATIFYCATION

(by academic year)

75,000

Surveys

14,000

Comments



**350** (2.5%)

Issues



# D2 NATURAL LANGUAGE PROCESSING

#### **TOKENIZATION**

Segmentation of text into words, punctuations and stop words.



Subfield of Artificial Intelligence which studies how computers process and analyze large amounts of natural language data.

#### **PART-OF-SPEECH**

Deduction of word's type, genre, number, etc. Also know as 'Linguistic annotations'

Token	No	he	tratado	con	él
POS	Adv	Aux	Verb	Adp	Pron
Tags	Neg	Sing 1Pers present	Past, Part	Prep	Masc Sing 3Pers

#### **SENTENCE DETECTION**

Finding and segmenting individual sentences

She is a great teacher that works a lot.

But he arrives allways late.

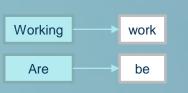
She is a great teacher

that works a lot

But he arrives allways late

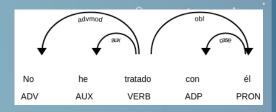
#### **LEMMATIZATION**

Extraction of the base form from a word.



#### **DEPENDENCY PARSING**

Detection of the relation between words in the sentence.



## NATURAL LANGUAGE PROCESSING

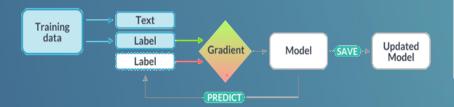




#### **TEXT CLASSIFICATION**

Process of categorizing text into different groups using a set of sample text labeled by a user and using an statistical model.

- Estimation of weight values base on examples.
- Comparation between estimation and labels.
- Measure of gradient to valuate result.
- Iterative re-estimation of weights.



#### **RULE-BASED MATCHING**

Process of defining patterns based on the contents and grammar of the sentence to identify text that match with defined patterns. It's based in Part-of-speech and Dependency Parsing and it's also called as Pattern Matching.

#### **Regular expressions**

Lower case of word contains 'profes'

#### **Words position**

Sentence starts with word of pattern 'A'

#### Use of lemmas

Lemma of word is 'teach'

#### Use of grammar

The first word it's a noun followed by an adjective

#### Use of dependencies

Subject of sentence is 'profe' and verb is 'teach' and object is 'student'

### 02 NATURAL LANGUAGE PROCESSING

**Use cases:** 



Speech **Word sense** recognition disambiguation





Natural language generation



Sentiment analysis

**Tools or libraries:** 







spaCy

**Natural Language** Toolkit (NLTK)

**TextBlob** 

**Stanford Core NLP** 

**Allen NLP** 

spaCy - Industrial-**Strength NLP** 

### 03 **METHODOLOGY AND SCHEDULING**

Project starts Data collection Working methodology Spacy initial explor.

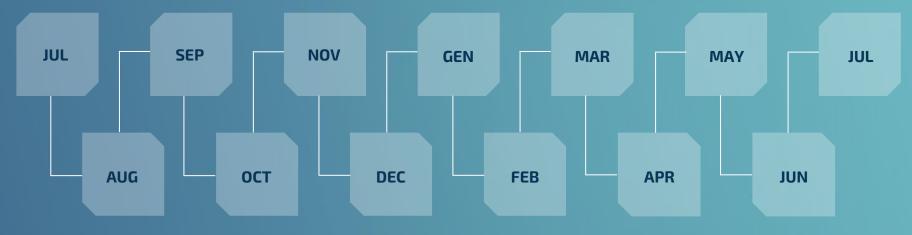
Iteration 1 User interf. validat. Django exploration Database definition

Iteration 2 Spacy deep explor. Adapt data to NLP Load data inJupyter

Iteration 4 NLP Spacy **TextCategorizer**  Iteration 6 and 7

Iteration 9

Master's Final **Project** presentation



Iteration 1 Data transformation User interface def. Spacy advan. explor.

Iteration 1 User interface Database implem.

Iteration 3 Language detection and enhancement

Iteration 5 **NLP Spacy** model 1 with Part-of-Speach Iteration 8

Iteration 10 Spacy and Django

Iteration 1. User interface definition

**Iteration 2**. Load of sample files and language detection

**Iteration 3.** Analysis and improvement of language detection

Iteration 4. NLP Spacy model 1. Issue 'No ha impartit clase'. Option 'TextCategorizer'

Iteration 5. NLP Spacy model 1. Option 'Parts-of-speech'

**Iteration 6.** User interface implementation

Iteration 7. Import surveys and comments

Iteration 8. Integration of Spacy model 1 in Django user interface

Iteration 9. NLP Spacy model 2. Issue 'Comentari problemàtic'

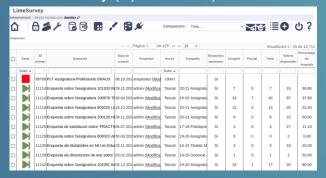
Iteration 10. Integration of Spacy model 2 in Django user interface

**Iteration 1. User interface definition**Original system

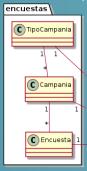
#### Gestor campañas (OCU)

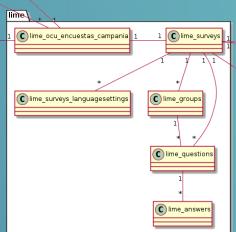


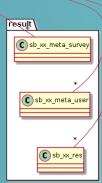
#### Lime Survey (Open source)



#### Database (MySQL)







# DEVELOPMENT Iteration 1. User interface definition Behaviour Driven Development (BDD)

#### List of features:

Authenticate

List campaigns

List survey comments

View survey comment

Edit comment

Import surveys

Process comments

#### Example of feature

Feature: List Campaigns

In order to list the campaigns imported from Lime

As a user

I want to see a list of all imported campaigns

#### Example of scenario

```
Background: There is a registered user

Given Exists a user "user" with password "password"

Scenario: Select a campaign from a list

Given I login as user "user" with password "password"

And A set of campaigns had been imported from Lime

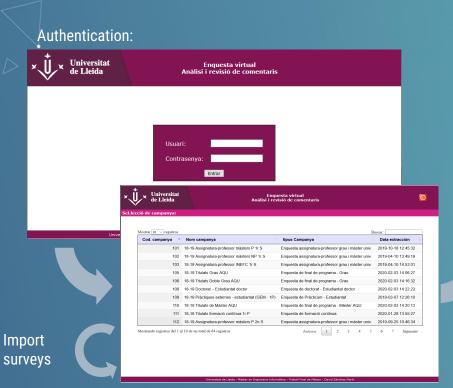
| cod_campania_lime | fecha_extraccion_lime | name | type_campaign_id |
| 101 | 2019-10-18 | 18-19 Assignatura-professor masters P 1r S | 25 |

When I list the campaigns

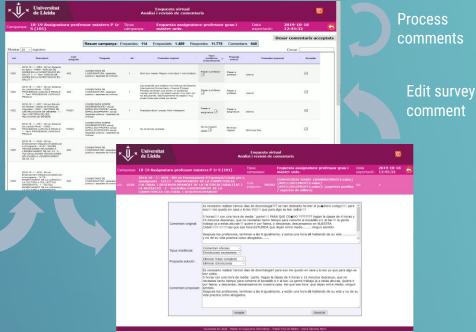
Then I'm viewing a list of campaigns

| Codi campanya | Nom campanya | Codi tipus campanya | Data extacció |
| 168 | 19-20 Practiques externes-estudiantat (GEM-1P) | 9 | 2020-11-17 |
```

# 04 DEVELOPMENT Iteration 1. User interface definition



List survey comments

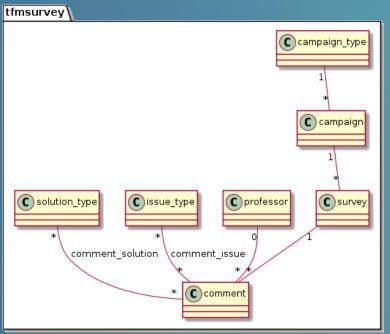


List campaigns

View survey comment

# 04 DEVELOPMENT Iteration 1. User interface definition

#### Database (MySQL)



#### Issue types

- The professor has not taught this group
- Problematic comments
- Offensive comments
- Professor comment on subject question
- Subject comment in professor question
- Spelling mistakes
- Exclamation points and excessive emoticons

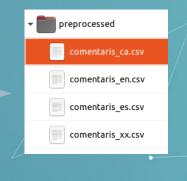
Iteration 2 and 3. Load sample files and language detection
Implementation



Load\_data.ipynb
(Jupyter Notebook)

LanguageDetector
PyCld2

Python 3.8 Jupyter 6.1.4 Pandas 1.1.3 Spacy 2.3 PyCld2 0.41



## **Iteration 2 and 3. Load sample files and language detection**Results

#### Comments by year:

Academic year	Total	Percentage
2017-18	90	1,14%
2018-19	1,468	18,56%
2019-20	6,353	80,31%
Total	7,911	

#### Comments by issue type:

Issue type	Total	%
Comentari problemàtic	271	53.88%
No ha impartit classe a aquest grup	115	22.86%
Comentari de professor	53	10.54%
Faltes d'ortografia	18	3.58%
Canvi de professor	17	3.38%
Comentari ofensiu	11	2.19%
Comentari d'assignatura	7	1.39%
Duplicar a professor	5	0.99%
Comentari excel·lent	4	0.80%
Exclamacions o emoticones excessius	2	0.40%
Total	503	



## Iteration 2 and 3. Load sample files and language detection Language enhancement

#### Language libraries comparative

	Idioma PyCld2				
Idioma Spacy	ca	es	en	xx	Total
ca	5,815	11	141	31	5,998
es	16	1,297	19	13	1,345
en	60	8	239	25	324
xx	68	28	30	55	181
Total	5,959	1,336	429	124	7,848

Divergence: 497 6,3%

#### Decision table

	Idioma PyCld2			
Idioma Spacy	ca	es	en	XX
ca	ca	ca	ca	ca
es	ca	es	es	es
en	ca	es	en	ca
XX	ca	es	са	ca/es

#### Comment by language

са	es	en	XX
6,197	1,365	239	55
78.9%	17.4%	3.0%	0.7%

# O4 DEVELOPMENT Iteration 4. NLP Spacy model 1. Issue 'No ha impartit clase' - Option 'Text classification' Load data

- Only models to Catalan and Spanish languages
- Division of comments in training and test sets: 80% / 20%
- > Division of sets in positive and negative cases

- Only questions about professors
- ➤ Inclusion of all the positive cases
- ➤ Selection of a subset of negative cases



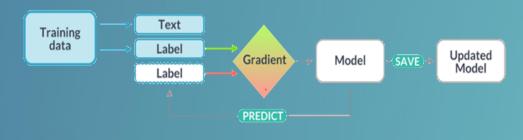
Iteration 4. NLP Spacy model 1. Issue 'No ha impartit clase' - Option 'Text classification'

Train model

- Limited training data
- ➤ Loop until Gradient < 0.001

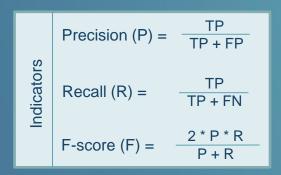
- > Creation of subsets of data (batches):
  - ✓ Compounding
  - ✓ Minibatch
  - ✓ Shuffle

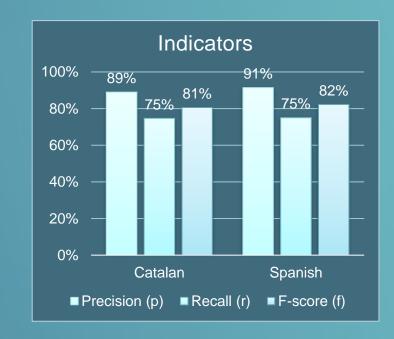




# O4 DEVELOPMENT NLP Spacy model 1. Issue 'No ha impartit clase' - Option 'Text classification' Evaluate

Cor	nfusion	User label		
table:		Positive	Negative	
Prediction	Positive	True positives (TP)	False Positives (FP)	
Predi	Negative	False negatives (FN)	True negatives (TN)	







model\_1\_no\_classes\_part-of-speech\_train.ipynb model\_1\_no\_classes\_part-of-speech\_evaluate.ipynb

Matcher

Based on Part-of-speech and Dependency parser

#### Transitive verbs:

- Starts with negative adverb
- Contain a verb from the verbs\_trans list
- Follows a noun from the noms trans list

#### Non-transitive verbs:

- Starts with negative adverb
- Contain a verb form the verbs list

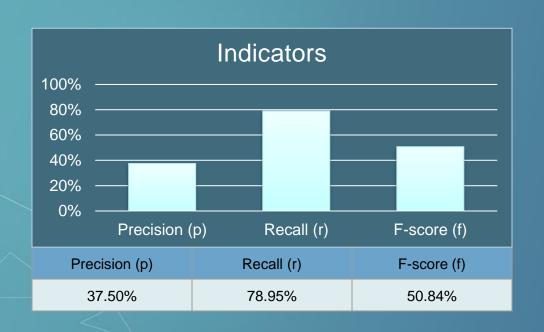
#### Rule for transitive verb:

```
adv_negs = ["no","tampoco"]
verbs = ["tratar","ir","conocer","ver","saber","ser","aparecer","tener"]
verbs_trans = ["realizar","dar","recibir","dictar"]
noms_trans = ["clase","asignatura"]
```



DEVELOPMENT
Iteration 5. NLP Spacy model 1
Option 'Rule-based matching'

04







- Long comments with more than one sentence
- Don't follow a clear pattern or regular expressions
- Only Catalan and Spanish languages

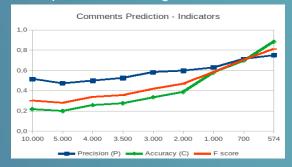
#### Options:

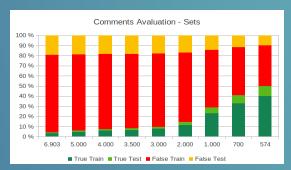
- Strategy: Full comments vs Individual sentences
- Sets size: Training vs test and Positive vs negative
- Pre-trained model: Large or small models

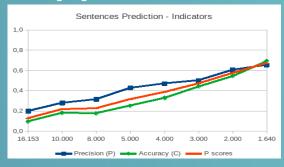
#### Conclusion:

- Full comments
- 700 comments (40% positives)

Comparative of strategies and set sizes. Catalan languages



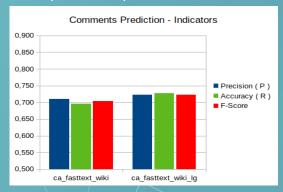






# 04 DEVELOPMENT Iteration 9. NLP Spacy model 2 Issue 'Comentari problemàtic'

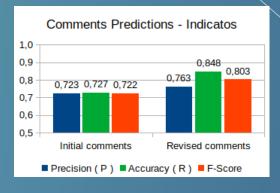
#### Comparative of pre-trained models

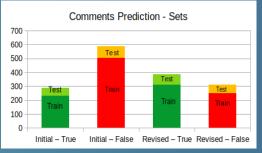


#### False positives problem:

- Most of the False positives could be True positives
- Reclassification of False positives comments:
   100 new positives
- Results:
  Accuracy increase 8%
  F-score: 12%

#### Results of comments revision





# DEVELOPMENT Iteration 9. NLP Spacy model 2 Issue 'Comentari problemàtic' Spanish language

Only 53 positives comments

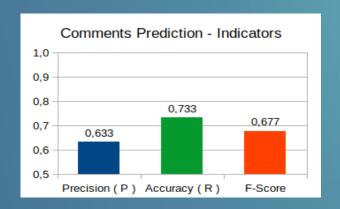
#### Parameters:

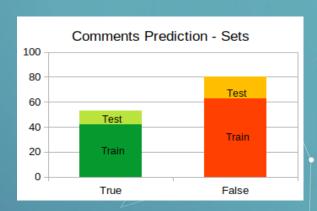
- Full comments analysis

- 40% positive cases

- Set size: 133 comments

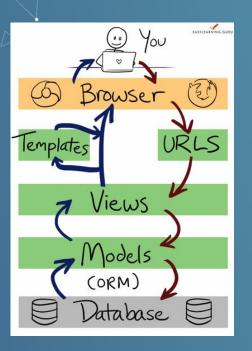
- Large pre-trained model





### Iteration 6. User interface implementation

### Django framework



Django is a high-level Python web framework that provide a lot of infrastructure and tools to help in the development.

His basic components are:

- Models: Definition of the objects and properties corresponding to the tables and

fields of the database.

Automate the management between the application and the database

- URLs: External addresses accessed by the user.

This component redirects to the real pages and manage the parameters

- Views: Responsible of executing the logic of the application. It obtains the

information from Models, redirect to URLs and send results to them.

Templates: Html pages with the format of the pages. They will be mixed with the information from views and can include other templates.

# 04 DEVELOPMENT Iteration 6. User interface implementation Models

#### diango [tfmsurvevs] ~/TreballFinalMaster tfmsurveysapp migrations > spacy > iiii static templates tfmsurveysapp base.html acampaigns list.html comment detail.html comments list.html display progress.html # form.html init .py admin.pv 🖔 apps.py forms.pv models.pv router.py tasks.py tests.pv 🖧 urls.pv views.pv manage.pv Pipfile Pipfile.lock

# README.md

#### /tfmsurveysapp/models.py

```
class CampaignType(models.Model):
    ...
class Campaign(models.Model):
    ...
class Survey(models.Model):
    ...
class Professor(models.Model):
    ...
class IssueType(models.Model):
```

```
class SolutionType(models.Model):
    ...
class Comment(models.Model):
    ...
class CommentIssue(models.Model):
    ...
class CommentSolution(models.Model):
    ...
```

```
class Campaign(models.Model):
    type_campaign = models.ForeignKey(CampaignType, on_delete=models.CASCADE)
    cod_campania_lime = models.IntegerField("Codi campania Lime")
    fecha_extraccion_lime = models.DateField("Data extracció Lime")
    name = models.CharField("Nom", max_length=50)
    import_date = models.DateField("Data importació")
    estat = models.CharField("Estat", max_length=50, null=True)
    subestat = models.CharField("Subestat", max_length=50, null=True)
```

# 04 DEVELOPMENT Iteration 6. User interface implementation URLs

#### django [tfmsurveys] ~/TreballFinalMaster tfmsurveysapp migrations > macy > Image static templates ✓ Implementation of the state of the st base.html acampaigns list.html acomment\_detail.html acomments list.html # display progress.html # form.html init .py 👼 admin.py 🖔 apps.py forms.py models.pv router.py tasks.py tests.pv 🖔 urls.pv views.py 🐌 manage.py Pipfile

Pipfile.lock

# README.md

#### Examples of urls

#### List campaigns:

http://localhost/tfmsurveys/campaigns/

#### **List survey comments:**

http://localhost/tfmsurveys/campaigns/22/
Campaign code=22

#### **View survey comment:**

http://localhost/tfmsurveys/campaigns/22/23405/ Campaign code=22, Comment code=22

#### /tfmsurveysapp/urls.py

#### diango [tfmsurvevs] ~/TreballFinalMaster. tfmsurveysapp migrations > m spacy > iiii static templates ✓ Image: ✓ Image base.html campaigns list.html comment detail.html acomments list.html # display progress.html # form.html init .py admin.pv 👼 apps.py forms.py models.pv router.py tasks.py tests.pv

🖔 urls.pv

manage.py

# README.md

views.py

# 04 DEVELOPMENT Iteration 6. User interface implementation Views

View's example 1

```
def campaigns_list(request):
    campaigns_list=Campaign.objects.all()
    context = {'campaigns_list': campaigns_list}
    return render(request, 'tfmsurveysapp/campaigns_list.html',
context)
```

#### View's example 2

#### /tfmsurveysapp/urls.py

```
Def campaign_list(request):
    ...
class CommentsList(ListView):
    ...
class CommentDetail(UpdateView)
    ...
```

# 04 DEVELOPMENT Iteration 6. User interface implementation Templates

Example of template: /tfmsurveysapp/templates/tfmsurveysapp/campaigns\_list.html

```
{% extends "tfmsurveysapp/base.html" %}
{% block content %}
{% load static %}
{% for campaign in campaigns_list %}
  >
     {{ campaign.cod campania lime }}
     {{ campaign.name }}
     {{ campaign.type campaign id }}
     {{ campaign.type_campaign.name }}
     {{ campaign.fecha_extraccion_lime|date:'d-m-Y' }}
     {{ campaign.import date|default if none:" "|date:'d-m-Y' }}
     <a href="{% url 'tfmsurveysapp:comments list' campaign.cod campania lime %}" target=" self">
          Veure
        </a>
  {% endfor %}
{% endblock %}
```

```
django [tfmsurveys] ~/TreballFinalMaster
tfmsurveysapp
migrations
> spacy
> 🗎 static

∨ I templates

   tfmsurveysapp
       base.html
       acampaigns list.html
       acomment detail.html
      acomments list.html
      # display progress.html
       # form.html
  init .py
   admin.pv
   👼 apps.py
   forms.py
   models.pv
   🏂 router.py
   tasks.py
  tests.pv
   🖧 urls.pv
  🖔 views.pv
manage.pv
Pipfile
```

Pipfile.lock

# README.md

Iteration 7. Import surveys and comments Multiple legacy databases access

diango [tfmsurvevs] ~/TreballFinalMaster

encuestas > teatures

> > spacy > = static ∨ I templates

> > tfmsurvevsapp

# base.html campaigns list.html acomment detail.html acomments list.html # display progress.html

# form.html 🎼 init .py

🖟 admin.py apps.pv forms.pv

umodels.pv

Fouter.pv

🏂 tasks.pv 🖐 tests.py

🐌 urls.py

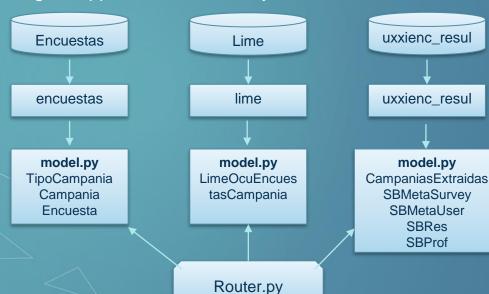
🚣 views.pv

uxxienc resul 🐌 manage.py Pipfile Pipfile.lock

# README.md

□ lime > lemplates tfmsurveys 🖿 tfmsurveysapp > migrations

The original application LimeSurvey include 3 database schemas:



#### Note:

**SBRes** 

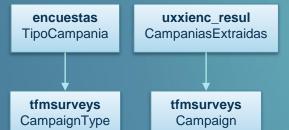
**SBProf** 

SBRes and SBProf: Raw queries and Multifield keys

# 04 DEVELOPMENT Iteration 7. Import surveys and comments

## Import Campaign types and Campaigns

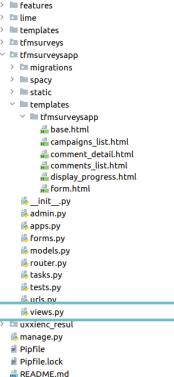
- Executed when opening Campaigns list
- Implemented in campaigns list view



#### **Import Surveys and Comments**

- Executed when the user press Import button in Campaigns list
- Implemented in ImportCampaign view
- Based on **RedirectView** standard view class





django [tfmsurveys] ~/TreballFinalMaster

encuestas

# 04 DEVELOPMENT Iteration 8 and 10. Integration of Django and Spacy

**Executing process** 

Showing results

l Irl·

http://localhost/tfmsurveys/campaigns/22/process/

New classes

C TfmLangDetector

init() detect(comment)

C TfmCategorizerModel1

init(language) test(comment)

C TfmCategorizerModel2

init(language) test(comment) Changes in views

**ImportCampaign** 

(RedirectView)

Import\_comments

**ProcessComments** 

(RedirectView)

Process\_models

**Celerys task** 

process\_model 1 process\_model 2

Urls:

http://localhost/tfmsurveys/campaigns/22/lang/ca
http://localhost/tfmsurveys/campaigns/22/issue/1

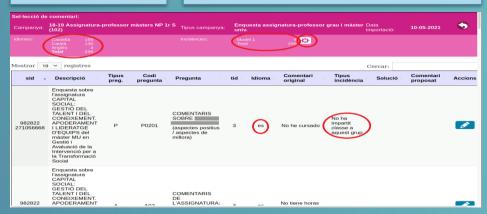
Changes in views

CommentList (ListView)

comments\_list.filter comments\_list.count Changes in template

comments\_list

- Process comments button
- Languages summary Issue types summary
  - Language column Issue type column



## 05 CONCLUSIONS

The aim of this Project was to develop an application that helps the Quality unit of the UdL manage the free comments included in the teaching surveys.

The main features of this application are:

- ► Importation of campaigns, surveys and comments from Lime Survey.
- List and search of campaigns, surveys and comments.
- Manual edition of comments.
- Processing of comments to identify their language and classify them in the issue types "Professor didn't teach this group" and "Problematic comments".



### 05 CONCLUSIONS



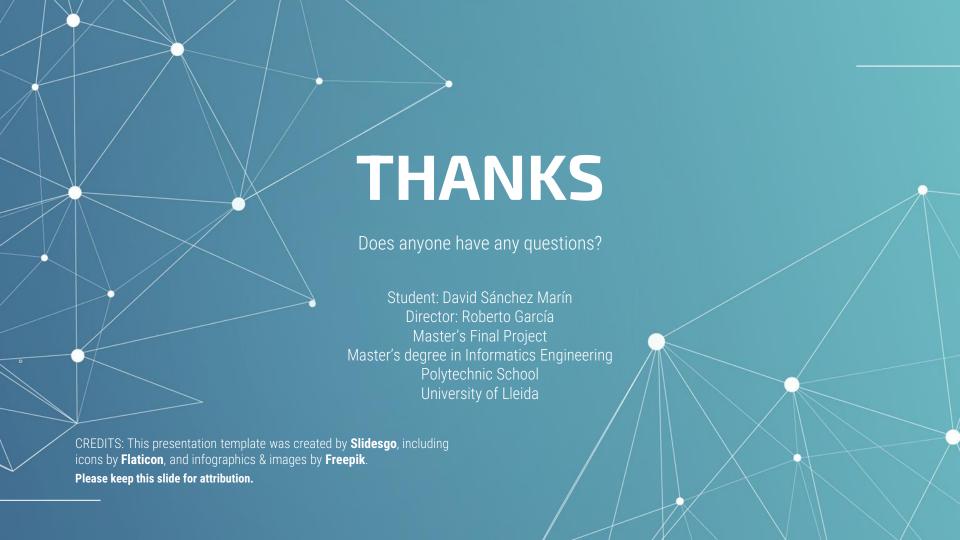
I defined the **features** for the high-level structure and **scenarios** for the detailed description.

I developed the user interface with **Python** and the **Django** framework.

I implemented its main components: **Models**, **Views**, **Urls** and **Templates**, but also access to legacy databases, virtual models, raw queries and others.

I analyzed and classified comments using the **Spacy** library.

I detected the language of comments with **LanguageDetector** and classified them in issue types with **TextCategorizer** and **Part-of-Speech**.



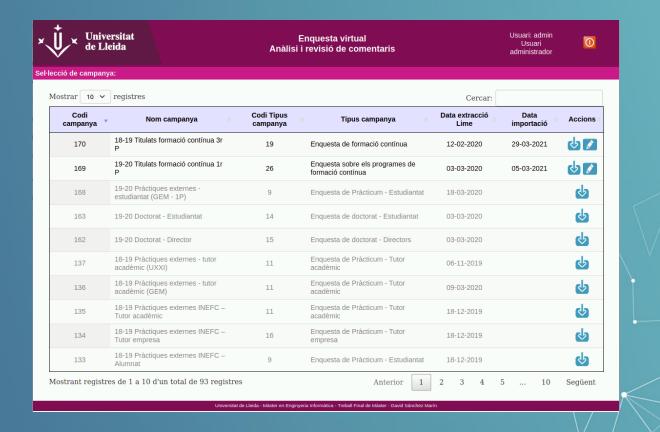


# **Annex. Screens**Authentication

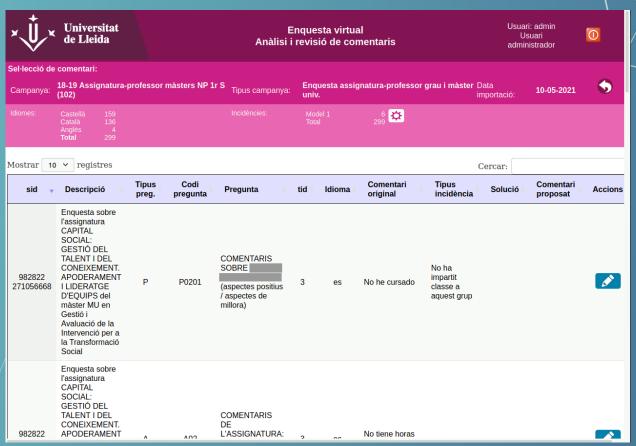


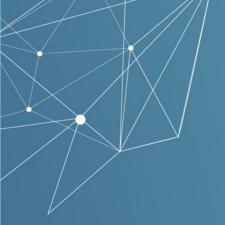
<b>▼</b> Universitat de Lleida		sta virtual ó de comentaris	
	Hamana at a lateral		
	Password:		
	Į.	ogin	
Universitat d	Lleida - Måster en Enginyeria Inforr	nàtica - Treball Final de Màster - David	Sánchez Marin

## **Annex. Screens**Campaigns list



## **Annex. Screens**Surveys and comments list





## Annex. Screens Edit comment

