

Document Categorization made easy: Start your own machine learning model (without coding)

Stefan Geißler, Kairntech

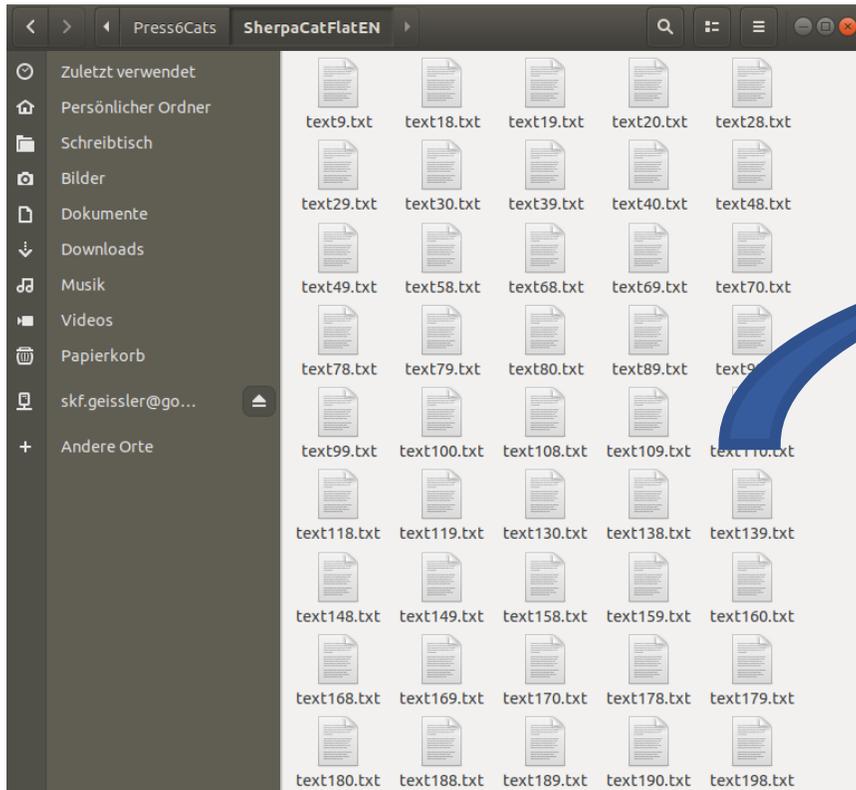


August 2020

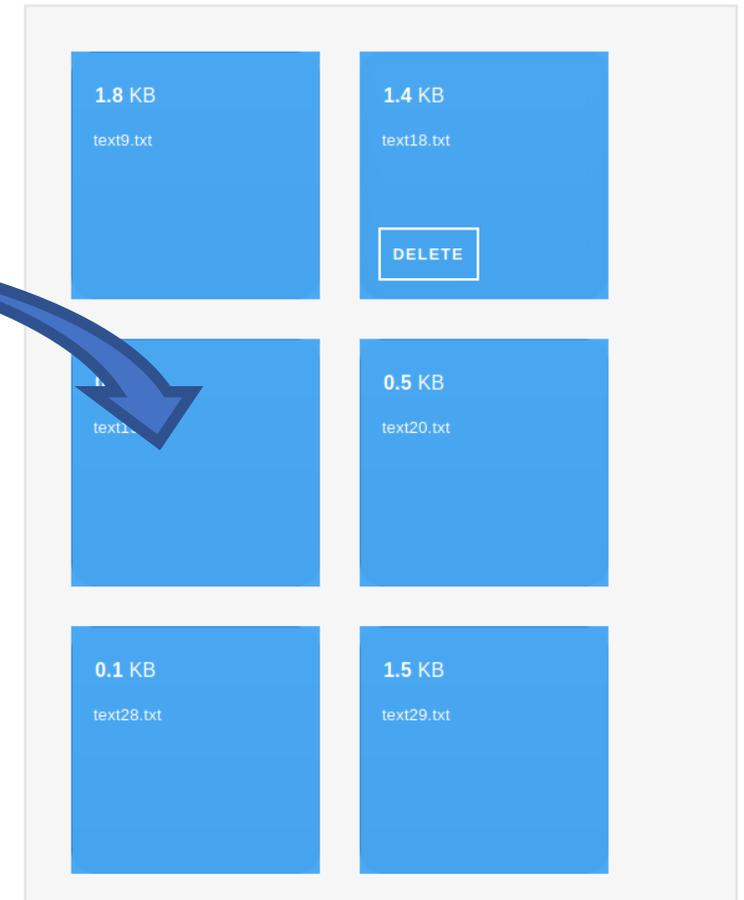
- ❖ **I need to sort a large number of documents into appropriate folders or categories**
- ❖ **I have no training corpus**
- ❖ **I don't want to (aehm: I don't know how to) program**
- ❖ **Still, I want to benefit from these new clever machine learning / AI / Deep Learning technologies that everyone is so excited about**

Solution: I do this in the Kairntech Sherpa!

Step 1: Import the documents



Import documents



Step 2: Define your categories

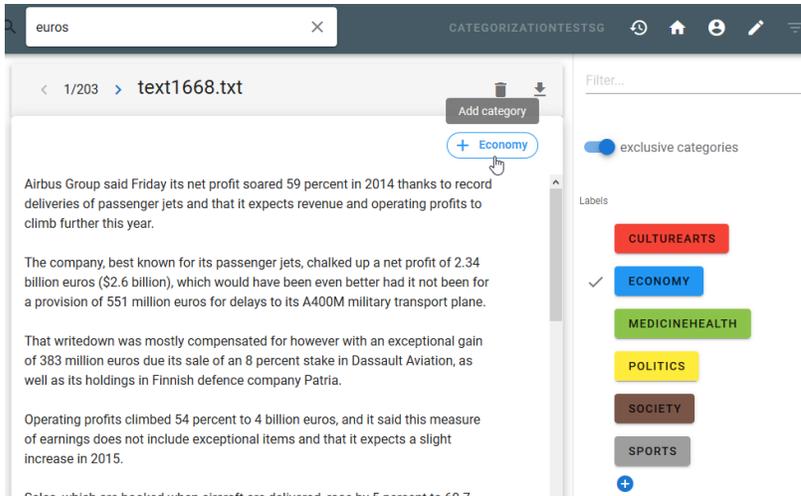
☰ **Kairntech Sherpa**

-  Project
-  Labels
-  Documents
-  Suggestions
-  Experiments
-  Test

-  Settings
-  Account
-  Jobs global view

-  CultureArts 
-  Economy 
-  MedicineHealth 
-  Politics 
-  Society 
-  Sports 

Step 3: Start putting documents into categories



euross

1/203 > text1668.txt

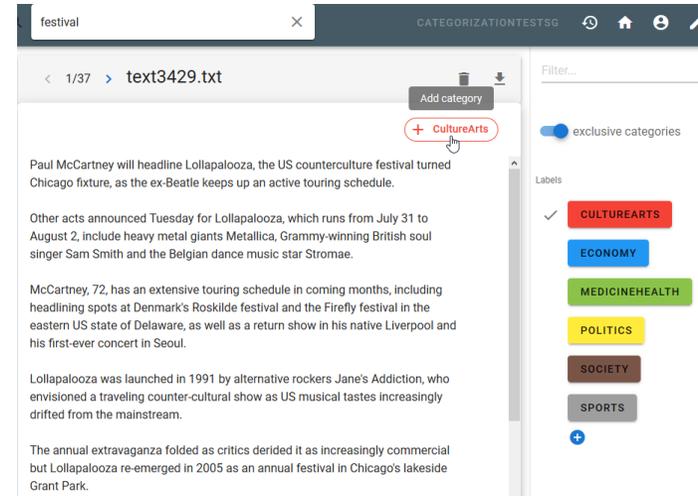
Filter...

exclusive categories

Labels

- CULTUREARTS
- ECONOMY
- MEDICINEHEALTH
- POLITICS
- SOCIETY
- SPORTS

Text content: Airbus Group said Friday its net profit soared 59 percent in 2014 thanks to record deliveries of passenger jets and that it expects revenue and operating profits to climb further this year.



festival

1/37 > text3429.txt

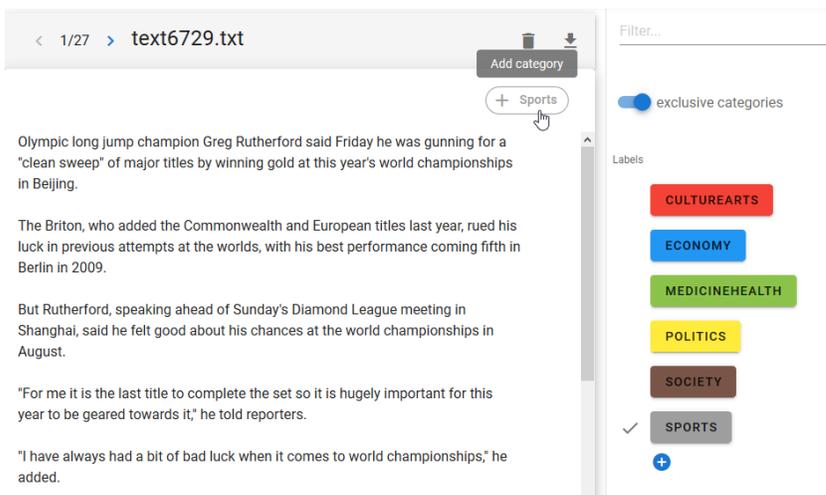
Filter...

exclusive categories

Labels

- CULTUREARTS
- ECONOMY
- MEDICINEHEALTH
- POLITICS
- SOCIETY
- SPORTS

Text content: Paul McCartney will headline Lollapalooza, the US counterculture festival turned Chicago fixture, as the ex-Beatle keeps up an active touring schedule.



1/27 > text6729.txt

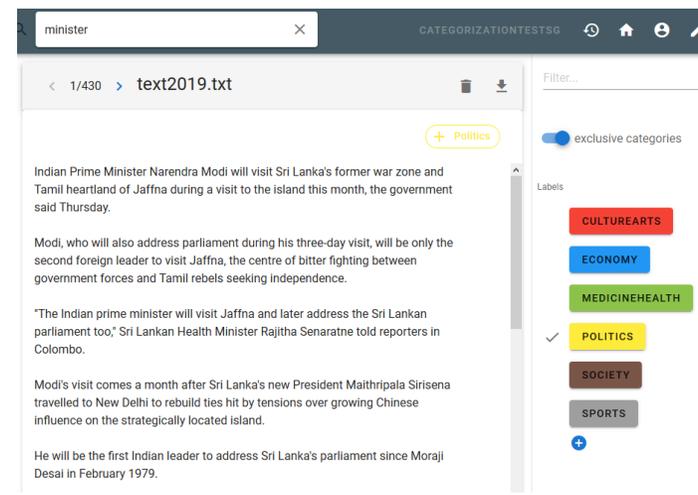
Filter...

exclusive categories

Labels

- CULTUREARTS
- ECONOMY
- MEDICINEHEALTH
- POLITICS
- SOCIETY
- SPORTS

Text content: Olympic long jump champion Greg Rutherford said Friday he was gunning for a "clean sweep" of major titles by winning gold at this year's world championships in Beijing.



minister

1/430 > text2019.txt

Filter...

exclusive categories

Labels

- CULTUREARTS
- ECONOMY
- MEDICINEHEALTH
- POLITICS
- SOCIETY
- SPORTS

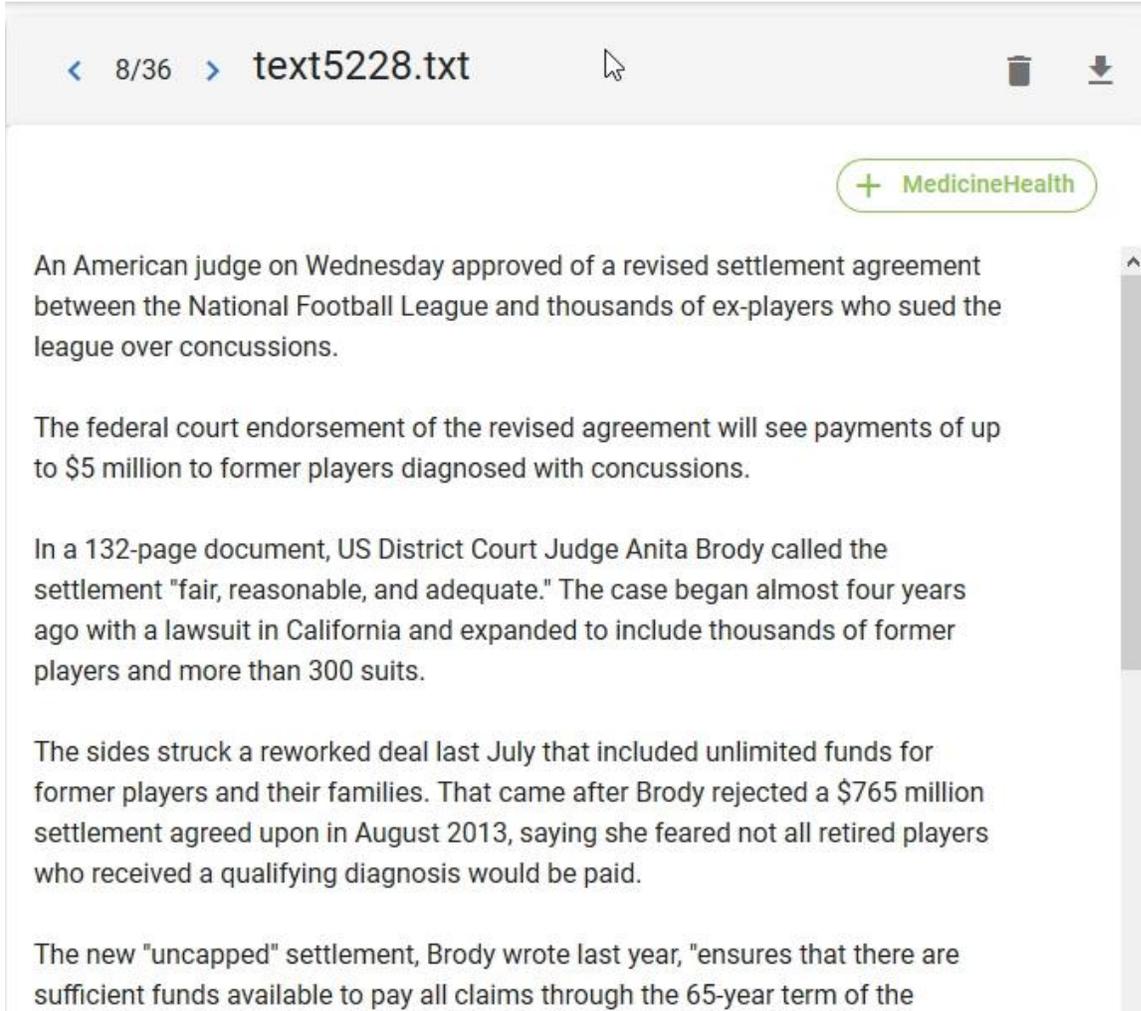
Text content: Indian Prime Minister Narendra Modi will visit Sri Lanka's former war zone and Tamil heartland of Jaffna during a visit to the island this month, the government said Thursday.

Here, my domain expertise is required. Assumption

- Many (not all) documents that speak about "championships" are about Sports
- "Minister" will indicate Politics.
- "Festival" CultureArts
- "Euros" may suggest Economy

I step through the search results for each of these and quickly collect 2-3 dozen sample documents for each.

Step 3: Unclear cases? Exclude!



< 8/36 > text5228.txt

+ MedicineHealth

An American judge on Wednesday approved of a revised settlement agreement between the National Football League and thousands of ex-players who sued the league over concussions.

The federal court endorsement of the revised agreement will see payments of up to \$5 million to former players diagnosed with concussions.

In a 132-page document, US District Court Judge Anita Brody called the settlement "fair, reasonable, and adequate." The case began almost four years ago with a lawsuit in California and expanded to include thousands of former players and more than 300 suits.

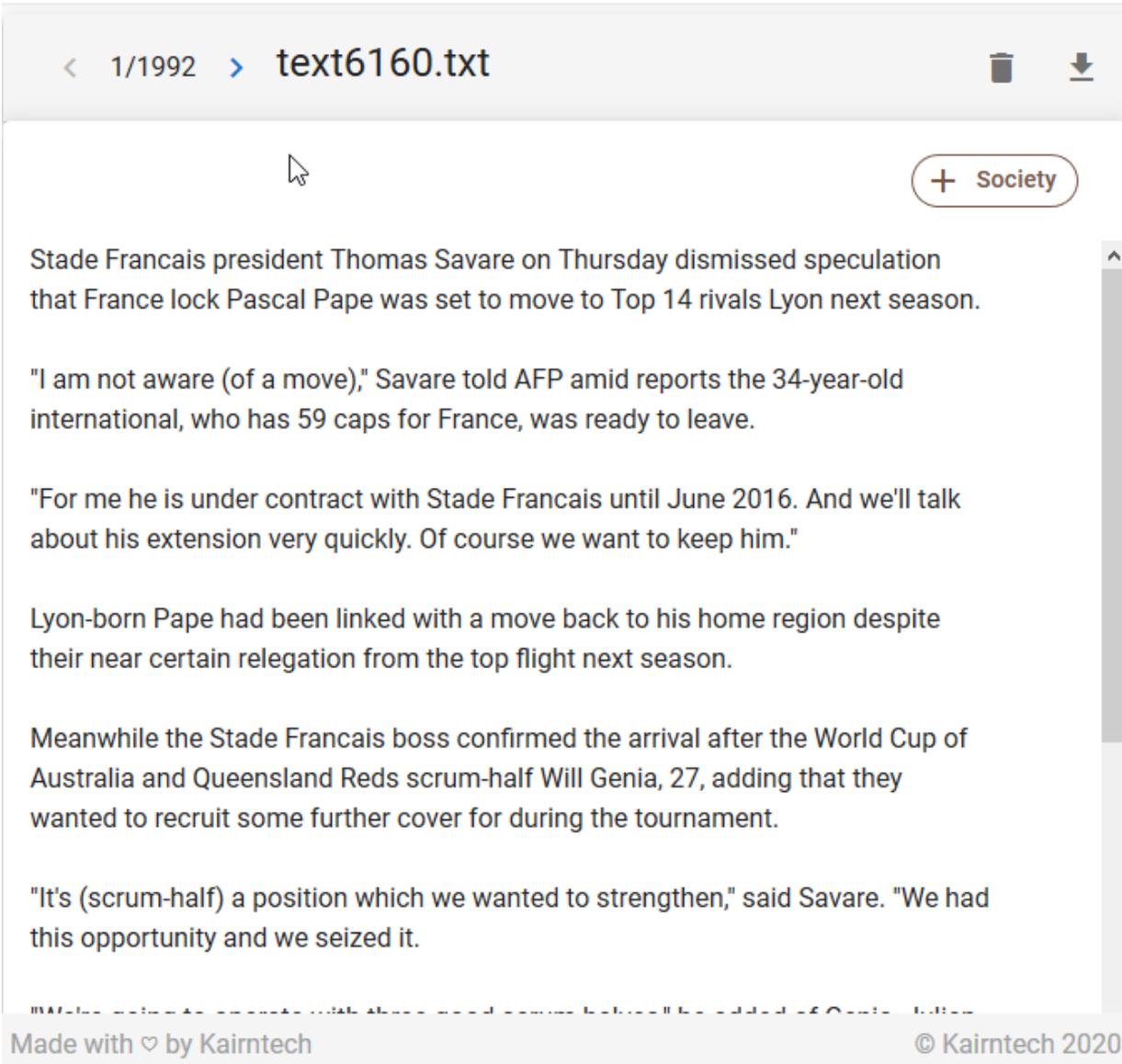
The sides struck a reworked deal last July that included unlimited funds for former players and their families. That came after Brody rejected a \$765 million settlement agreed upon in August 2013, saying she feared not all retired players who received a qualifying diagnosis would be paid.

The new "uncapped" settlement, Brody wrote last year, "ensures that there are sufficient funds available to pay all claims through the 65-year term of the

Text mentions the topics Sports and MedicineHealth issues and legal/economic topics ...

When unsure, just exclude the document for a start or assign it to more than one category.

Step 4: System starts to learn ...



< 1/1992 > text6160.txt  

 + Society

Stade Francais president Thomas Savare on Thursday dismissed speculation that France lock Pascal Pape was set to move to Top 14 rivals Lyon next season.

"I am not aware (of a move)," Savare told AFP amid reports the 34-year-old international, who has 59 caps for France, was ready to leave.

"For me he is under contract with Stade Francais until June 2016. And we'll talk about his extension very quickly. Of course we want to keep him."

Lyon-born Pape had been linked with a move back to his home region despite their near certain relegation from the top flight next season.

Meanwhile the Stade Francais boss confirmed the arrival after the World Cup of Australia and Queensland Reds scrum-half Will Genia, 27, adding that they wanted to recruit some further cover for during the tournament.

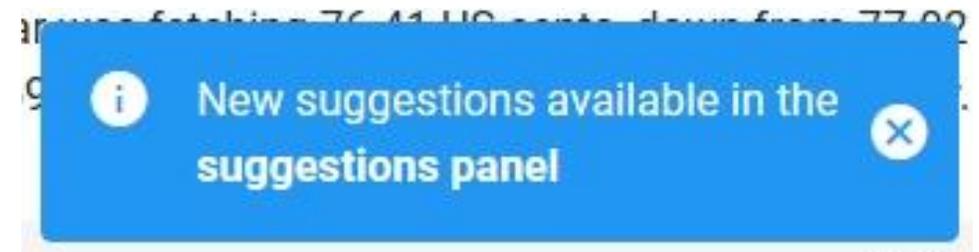
"It's (scrum-half) a position which we wanted to strengthen," said Savare. "We had this opportunity and we seized it.

"Make points to compete with those need some help" he added of Genia's fellow

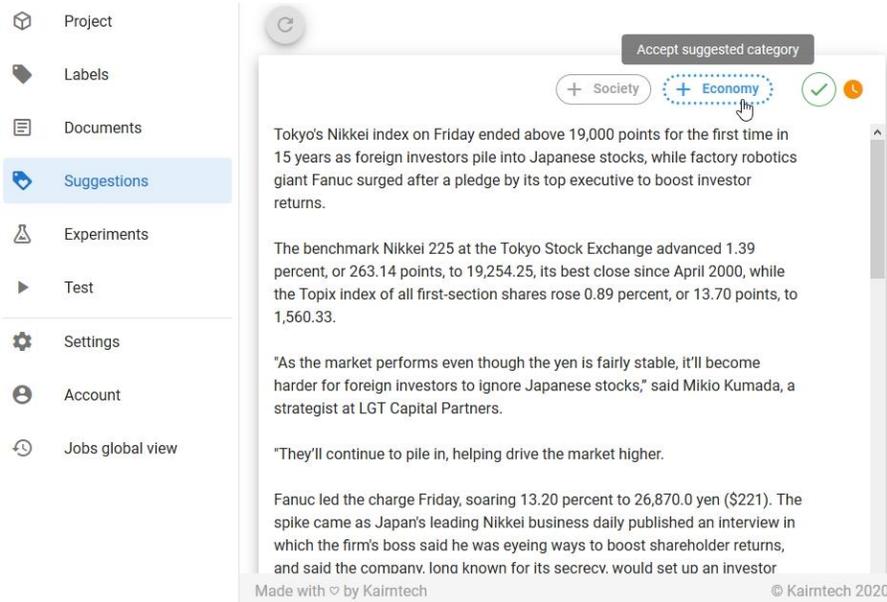
Made with  by Kairntech © Kairntech 2020

After a few minutes of assigning documents to categories, the system starts to learn.

... and generates automatically new suggestions.



Step 4: accepting suggestions (or not) ...



System presents documents it “thinks” are about a given category.

I accept and immediately the next suggestion pops up.

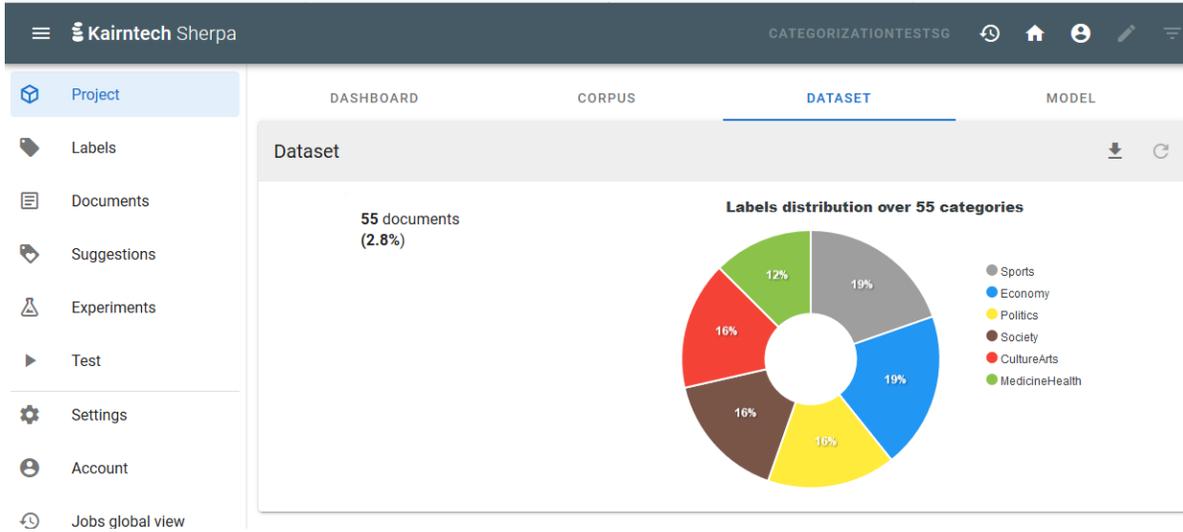
This is faster than having to select documents and their training categories by myself.

Validate document categories

Economy ×

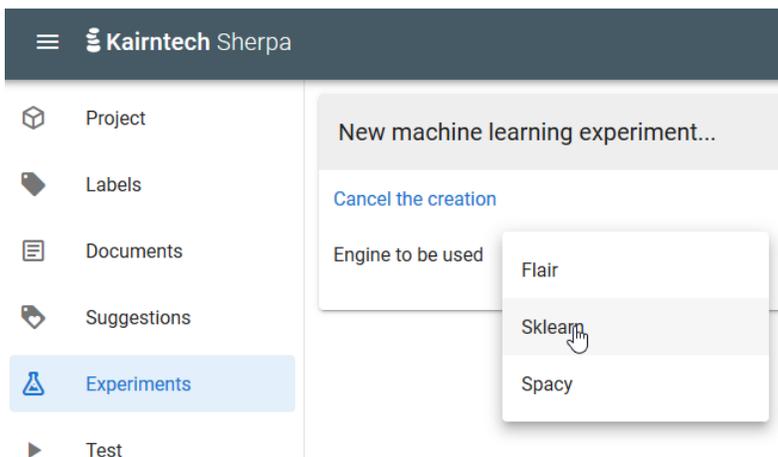


Step 5: where am I?

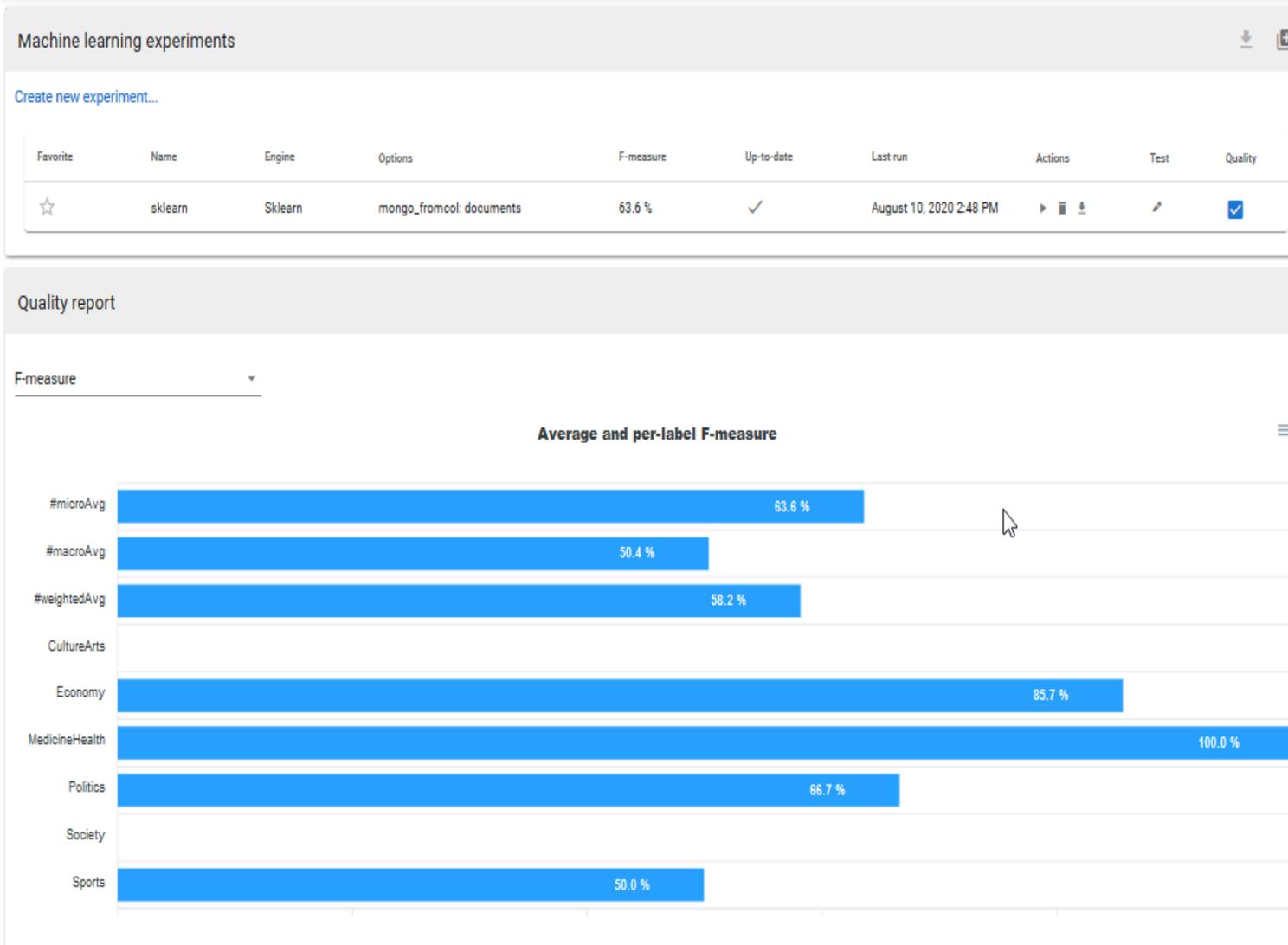


Check how many examples I have created?

Ok, time for a first evaluation!



Step 6: some orientation



Ok, some categories are already quite ok.

“CultureArts” and “Society” seem to still require some work.

Let’s go back to assigning some more in these categories ...

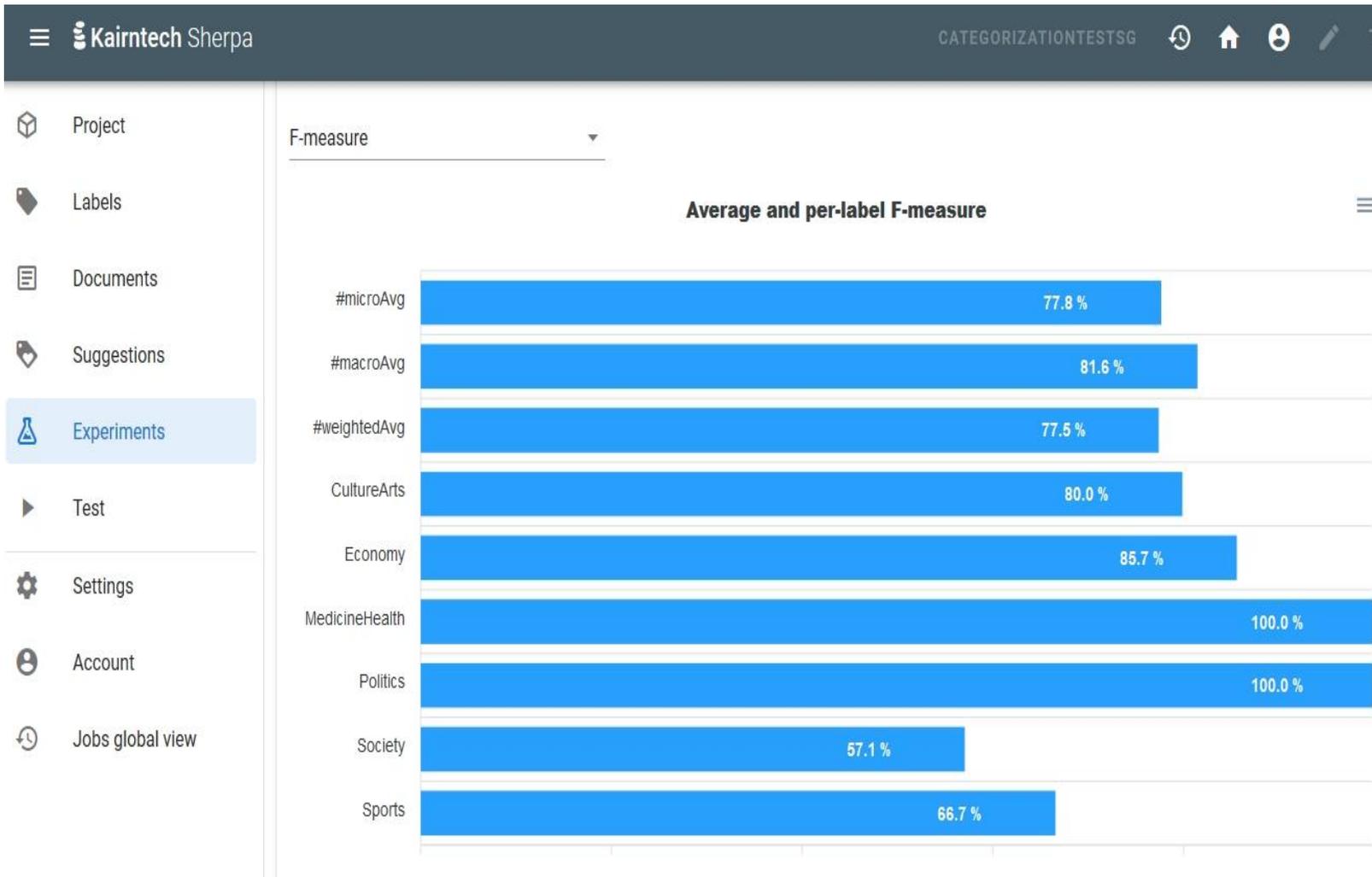
Step 7: ... after adding some CultureArt texts



Ah we are getting there!

But “Society” still seems to need some attention ...

Step 7: ... after adding some CultureArt texts



See!

Until here we have spent roughly 30-45 min.

We have started from scratch with uncategorized documents using only our domain knowledge

Arriving at ~78% quality.

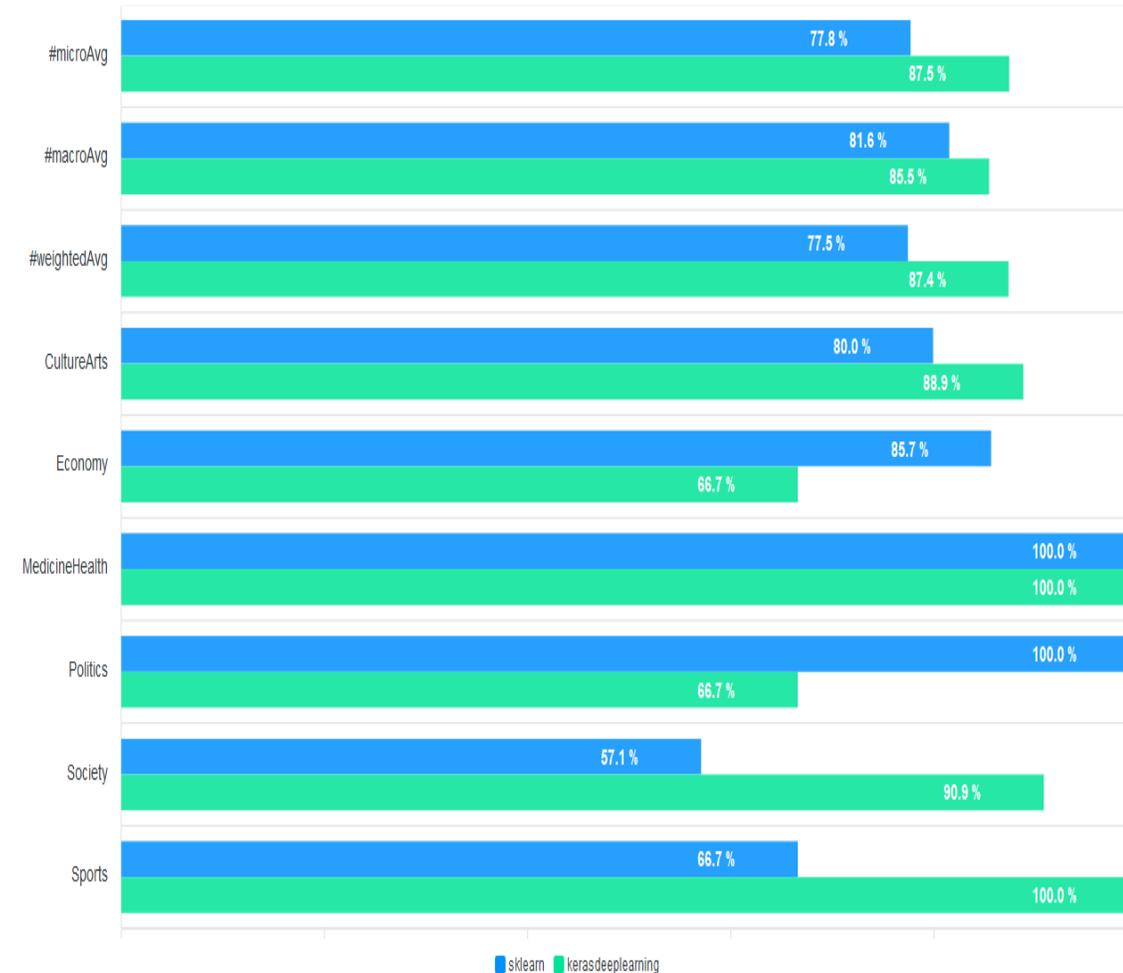
F-measure	Up-to-date	Last run
77.8 %	✓	August 10, 2020 4:00 PM

Step 8: Tuning

- ❖ So far we have « only » used the default categorization (quick, applicable to most scenarios, but not the top approach that Sherpa has to offer)
- ❖ Let's now apply a Deep Learning based approach – all available from the Web GUI

kerasdeeplearning	Sklearn	train_on: train, test_on: test, mongo_fromcol: documents, classifier: KerasMLPClassifier, featurizer: tfidf	87.5 %
-------------------	---------	---	--------

- ❖ The Deep Learning based approach boosts the quality from the default 78% up to 87% on out little manually created training corpus.
- ❖ Experimentation/training time sofar: Still less than 1 hour.



- ❖ **Easy intuitive handling of the application**
- ❖ **Instant, online learning, immediate feedback**
- ❖ **Useful results after very short time**
- ❖ **Need better results?**
 - ❖ **Invest some more time, adding manually labelled data**
 - ❖ **Experiment with other learning methods**
- ❖ **I now have a categorizer and I didn't have to bother anyone to implement it for me.**

**The Kairntech Sherpa:
Powerful Machine Learning for non-programmers.**

www.kairntech.com