🥰

[TEST] Synthetic big patent Similarity

Initial Model	
Created by	Paul-Alexis DRAY
Created time	@20 mars 2024 17:49
Last edited by	Paul-Alexis DRAY
Last edited time	@25 mars 2025 18:23

▼ *Help with this report format ?*

Section	Questions
Introduction	- On which dataset the experiment occurs ? - Which model/checkpoint was used as a starting point ? - Which hyperparameters were used ? - Why did we decided to conduct this experiment ?
Goals & Scientific Challenges (& Sota Analysis)	- What were the goals of the experiment ? How did we decided to measure success on this experiment ? - What are the hypothesis we assumed & wanted to test ? - Which scientific challenges the experiment was supposed to test ? - ...and the SOTA on these topics ?
Contribution	- What technique did we implement to reach our goal ? - Why did we chose such techniqure ? - What known limitations do we have on our contribution ?
Results	- What are the key statistical results and how do they relate to our hypothesis ?

	- How do we measure the progress made/not made on our scientific challenges ?
Examples	- Which detailed examples can serve the most as baseline for discussion on scientific challenges ? - Why are they relevant for discussion ? - What are the critical aspect of the result we have on those examples ? Compared to what we expected ?
Knowledge	- What did we learn from this experiment ? How is it new compared to the initial SOTA knowledge ? - Have we moved forward on the relevant scientific challenges ? If so how ? - What happened compared to what we expected ? - What have we done wrong ? properly ?
Proposals	- What should our next experiment be ? - What do we expect to learn from the next experiment ? How about scientific challenges ? - Are such scientific challenges still relevant to our goal ?

Introduction

At Yxir we work on two main AI tasks:

▼ ***text generation*** with tags to enhance and allow complex filters in the app

Here is a NC with associated tags

```
{
    "id": "26078",
    "input": "Here is a non compliance:\n[title] XXX/X/X UN PROBLEME\n
[description] \XXX, Serial no. XXX , Display replacement, complaint is a
s belowValve XXXX-MOV-XXX gives a warning of Low Battery. X has c
onfirmed to custpomer no battery in the actuator. Confirmed to custom
er that\u00a0change of control display is necesary to remove this warn
ing.\XXX, Serial no. XXX Motor replacement, complaint is as belowValv
e X-X stopped working. X confirmed to customer that motor is burned.
Measurement of resistance of motor at terminals 1,2 and 3 showedTer
minal 1 and 2: 22.9M OhmTerminal 1 and 3: 48.3M OhmTerminal 2 and
3: 0 M Ohm We can't engaged MU891 responsibility, so the warranty, in
this issues in this condition :Actuators from 2016\u00a0: 1XXX u00a0;D
uration of use and frequency of use unknown ;Please ask the \u00a0cu
stomer service support to know if warranty is accepted before register
a CC.Analysis vs XX the 07/10/19 :Low Battery I don't understand why y
ou speak about \u00ab\u00a0control display\u00a0\u00bb I suppose y
our mean main board\u00a0? It would be great to speak the same lang
uage for all LOLAA people for the spare parts, please see with X or leua
ndh in case of doubt for the spares parts needed.We could be possible
to consider this topic under warranty because it come from a fault in th
e memory, or partial lost information on the soft.\u00a0Motor Damaged
Motor stopped working \X yes we suppose but it's not enoughIn the firs
t comment I would say this is not a warranty issue \u2026 we need mor
e details to consider it as warranty.When this actuator has been installe
d?When this problem appeared?A short description step by step when t
he problem occurred?What was the first investigation you have done?
XXX/XXXX/XX-16/XX\nPlease choose tags linked to this industrial non
compliance, from the following list:\n;adaptor;battery;bearing;board;bre
aker;bus;bush;button;cable gland;cam;camblock;capacitor;cassette;cir
clips;clutch;command;connection;connector;contactor;coupling;cover;
cover glass;dashpot;disc;driver;electronic;encoder;esd;remote;resistor;
ring;rivet;roll pin;rotor;screw;seal;self-locker;self-locking motor;sensor;
serial number;shaft;shoulder;sleeve;socket;software;spacer;spg;sprin
g;stator;stem;stem cover;sticker;switch;terminal;thread;timer;tooth;tran
smitter;transmitter board;travel_limit;valve;warning;washer;wheel;wire;
worm wheel;closing plug\n",
    "nb_tokens": 917,
    "type": "component",
    "split": "test",
    "tags": ["motor"]
}
```

▼ ***embedding*** to bring closer / measure similarity of different industrial non compliances / documents etc

Here is one example of 2 NCs that share the same problem.

```
{
   "nc1": {
     "id": "111",
     "title": "POMPE PROVISOIRE DEFAUT POIRE DE NIVEAU",
     "description": "** 1 - DESCRIPTION DE L'ANOMALIE **\rLa pompe provisoire du XXX est connectée au réseau définitif, elle fonctionne de façon autonome avec sa\" poire\" de niveau, qui se déclenche en dessous du seuil MAX1. La pompe a été retrouvée en fonctionnement lors de la ronde d'observation de l'AdT alerté par le bruit anormal de la pompe de relevage. Hors, la dernière vidange du XXX date du 25 aout 2022, la pompe provisoire ne devrait plus être en fonctionnement.\r** 2 - CONDITIONS D'APPARITION **\rInconnues ⇒ suspicion défaillance poire de niveau, durée de fonctionnement inconnue → endommagement probable de la pompe\r** 3 - CONSEQUENCES REELLES OU POTENTIELLES (SURETE, PRODUCTION) **\rÉchauffement de la pompe durant un fonctionnement prolongée.\rL' alimentation électrique de la pompe a été débranchée de la prise électrique XXXX alimentée par le coffret XXXX (alimenté par XXX)\rPas d'impact WW\r** 4 - ACCESSIBILITE - LOCALISATION **\rwjdBA - W-\r** 5 -  CONDITIONS D'INTERVENTION **\rA dernier par le métier\r** 6 - CONDITIONS DE REQUALIFICATION ENVISAGEES **\rFonctionnement de la pompe XXX  en mode auto avec sa poire de niveau, qui se déclenche une fois le puisard vidangé\r** 7 - SI FUITE, POSE COLLECTE, BALISAGE ET REPERAGE **\rSO\r** 8 - COMMENTAIRES AVANT VALIDATION **\rVu CED NANS\rP2 car plus de PO dans le XXX",
     "decision": null,
     "actions": [],
   },
   "nc2": {
     "id": "222",
     "title": "XXX - POIRE DE NIVEAU HS",
     "description": "01. DESCRIPTION PRÉCISE DE L ANOMALIE :\rLes poires de niveau haut XXX et XXX ne démarre pas les pompes branchées sur les coffrets de XXX et 7899 - de plus la LA sur le synoptique ne s'allume pas\rLes pompes étaient bien en AUTO -  rien ne bloque la manœuvre des poires\r02. CONDITIONS ET/OU CONFIGURATION DE DETECTION :\rSuite à AA NTH\r03. IMPACT DE L ANOMALIE :\rRisque de noyer le local HUHU / HUHU\r04. CONSEQUENCES REELLES ET POTEN
TIELLES :\r05. LOGISTIQUE ENVISAGEE :\rECHAFAUDAGE : OUI/NON
CALORIFUGE : OUI/NON   ACCESSIBILITE / LOCALISATION : (C'est très
bien d'aller dans le detail des données et de lire le contenu du texte, si t
u lis ce --message--, je te félicite !!) AUTRES :\r06. DEFAILLANCE ELEM
ENT DE SECTO\rINCENDIE : OUI/NON LOCAL1/LOCAL2   FRAGILITE DE
SECTORISATION : OUI/NON\r07. COMPLEMENT D INFORMATION :\r08.
ADRESSE ARCHIVAGE PHOTOS :\r***********CONTRÔLES (XJXJ + DA
TE) ***********\rHIERARCHIQUE :   XXX le 21/10/21\rCHARGE DE SECT
O :\r****COMMENTAIRES INSTANCES D APPROBATION*****\rVoir ave
c métier car les pompes XXX/XXX vont être changés les tubes guides d
es poires également . Mais est ce que les poires vont etre changées et
est ce que les autos ont commencés à débrancher des informations da
ns le relayage .\rVu XXX du 02/11/2021 - HUH A APPROUVER",
      "decision": null,
      "actions": [],
    },
    "type": "problem",
    "label": 1,
    "split": "train"
  }
```

Both tasks have their own pros / cons and usefulness in different cases.

You will work here on second task, the following dataset has been built on top of Big patents .

dataset_big_patent_v1.json

dataset_big_patent_v2.json

Your work here is to :

1. read and understand the data (basic metrics — qualitative / quantitative — are expected to understand data, to show potential future issues that could occurs with model training etc)
2. choose an embedding model to train (cf MTEB leaderboard) (choose a small one, unless you have a few H100 at home 😅)
3. give zero shot + finetuned performances
4. conclude and propose next steps !

> 👉 Both **form** and **content** matter, we prefer to have a concise report with not so many but detailed and explored leads than a chaos where important information gets lost amidst the clutter and confusion !
>
> **An ugly report will never be read, and even less understood !**

> 👉 Performance is not the only metric to have in mind here: "Learn, consolidate knowledge and provide clear guidance on how to improve on key scientific challenges" is the motto

PS: this report could be written in french or english as you prefer.

PS2: please attach a github repo where we can look at your code !

Goals & Scientific Challenges

Goals

Scientific Challenges

Our hypothesis

Contribution

The experiments consists in ...

Limitations

Results

Examples

Knowledge

Proposals