Python的NLP实战分享

如何实现合同风险预测模型?

GVA TECH Co., Ltd
藤井美娜

自我介绍

藤井美娜 inazo18

GVA TECH

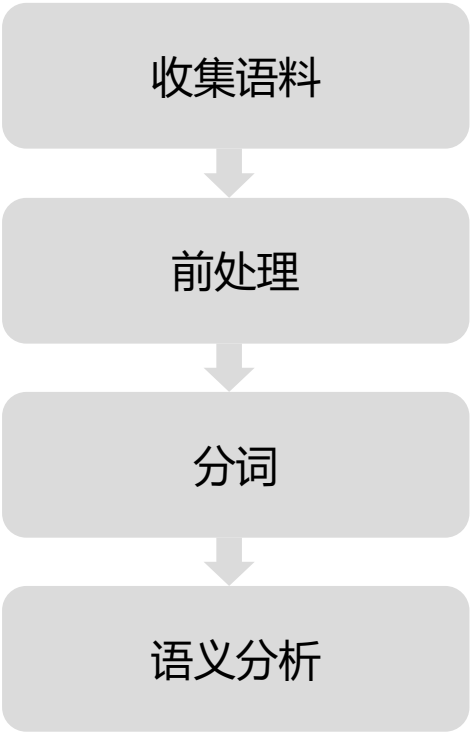
- Machine Learning Engineer / Data Scientist
- GVA TECH的人工智能法律服务AI-CON的多语言系统开发负责人

目录 CONTENTS

1 Python NLP 入门

简单介绍自然语言处理的流程和使用corpus的EDA方法。

NLP基础



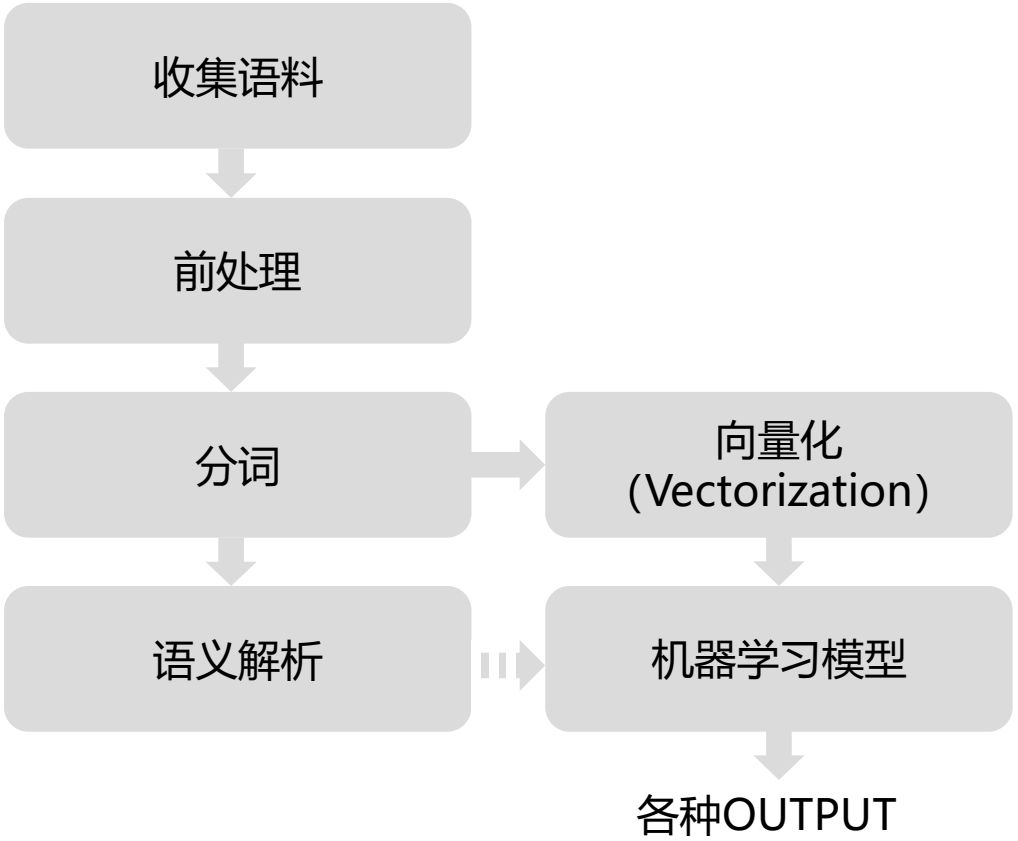
- 使用爬虫收集的语言数据
- 公司拥有的语言数据

- 根据你要达到的目的删掉一些不需要的东西，例如①②③；《 》等特殊符号。

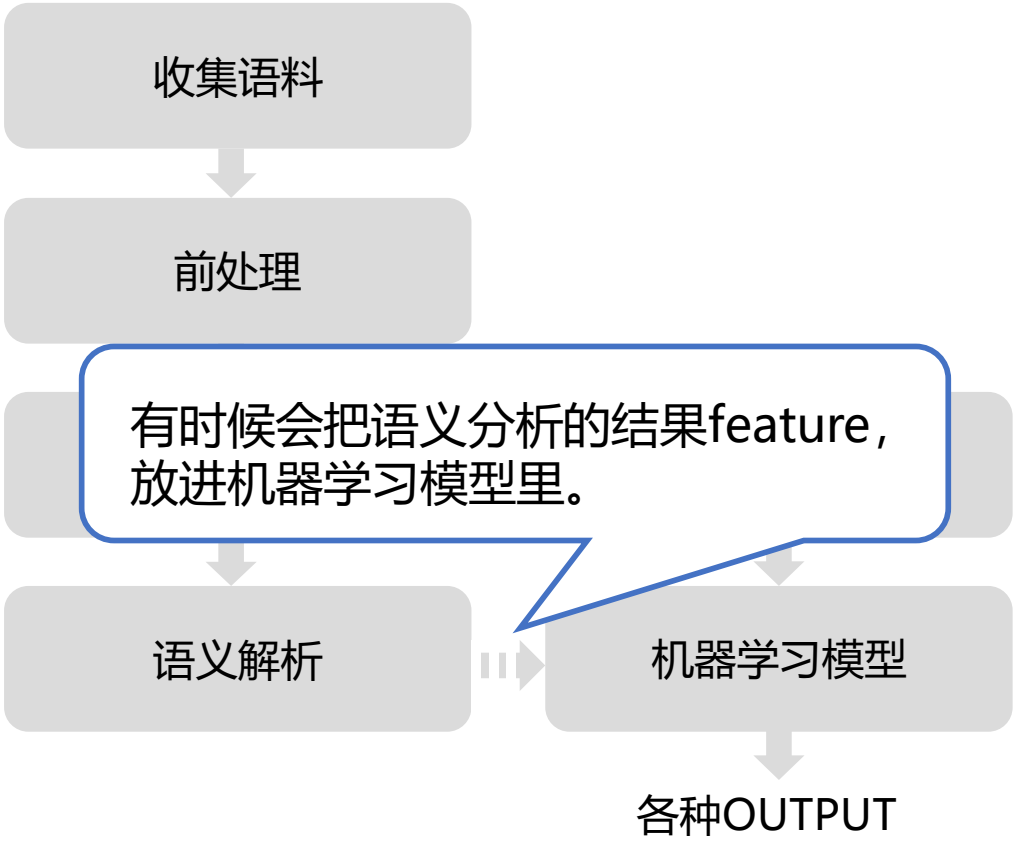
- 确认需不需要分隔词素

- 词语
- 短语
- 句子
- 文

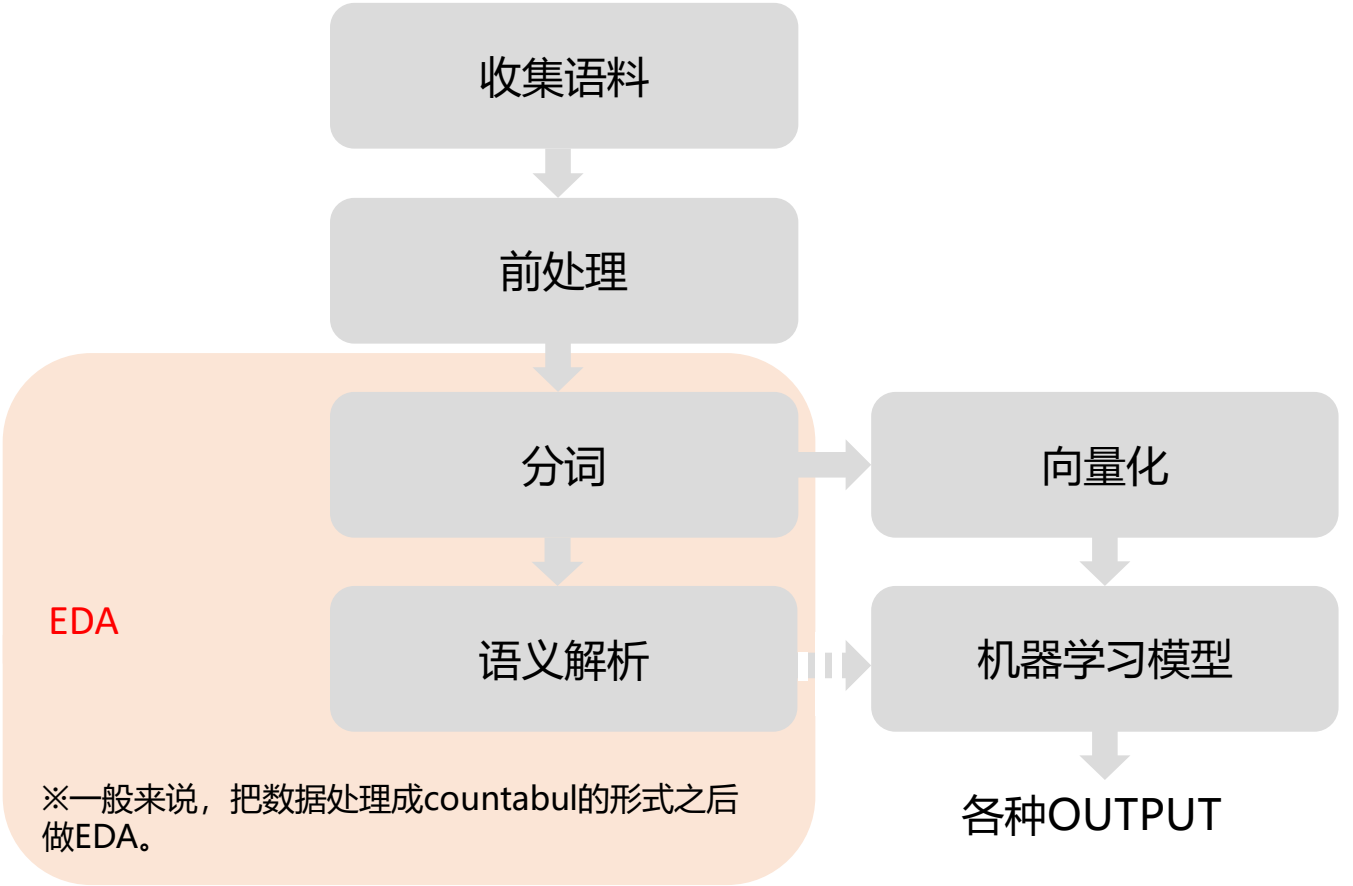
NLP基础



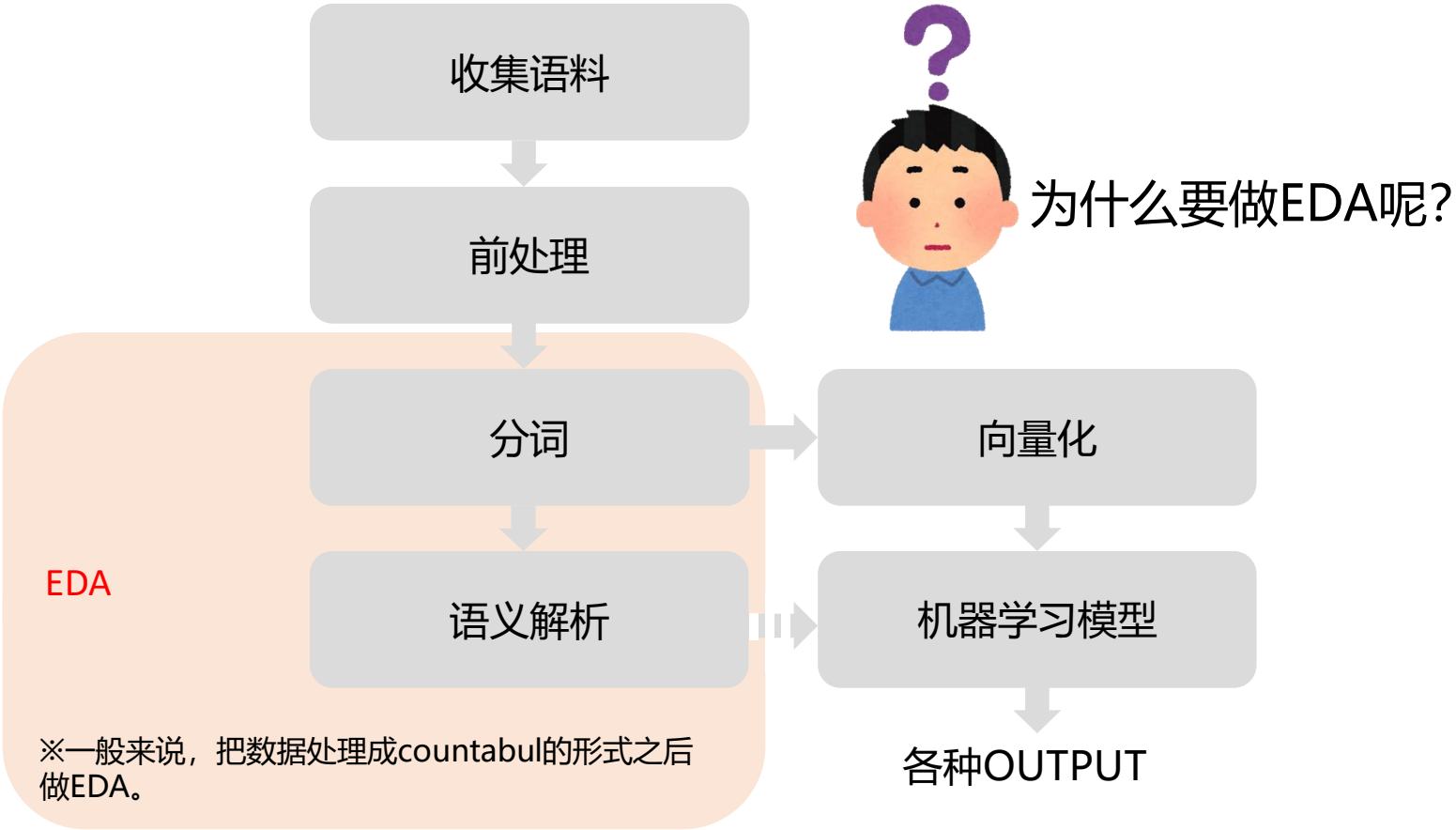
NLP基础


收集语料
前处理
有时候会把语义分析的结果feature，
放进机器学习模型里。
语义解析
机器学习模型
各种OUTPUT

NLP基础


收集语料
前处理
分词
语义解析
向量化
机器学习模型
各种OUTPUT
EDA
※一般来说，把数据处理成countabul的形式之后做EDA。

NLP基础



为什么要做EDA呢?

EDA

- EDA（Exploratory Data Analysis/探索性数据分析）
- 各个领域都具有那个领域的语言特征，我们不能不考虑。

例如：

“网络上的语言表达”
“文学作品上的语言表达”
“合同书上的语言表达”
“病历卡上的语言表达”

- EDA有很多办法，下面介绍使用corpus的词类分析方法：
  - 首先收集10万条合同文章，另外准备别的领域的比较稳定的corpus做来benchmark。
  - 使用同一个办法来处理各领域的语言数据做一个词类频次分布表，在各词性出现频次涂上颜色。

使用corpus做EDA

		合同文章		出版図書	出版雑誌	出版新聞	図書館書籍	政府報告	課文	宣传文章	畅销書籍	Yahoo!知识库	Yahoo!博客	韵文	法律文章
		GVA収集	BCCWJ全体	各レジスター（頻度1までの見出し）											
		契約書データ（補助記号を入れて	PM～OWは統合 OT～OMは可変	出版書籍	出版雑誌	出版新聞	図書館書籍	特定目的白書	特定目的教科書	特定目的広報紙	特定目的ベストセ	特定目的Yahoo!知	特定目的Yhaoo!ブ	特定目的韻文	特定目的法律
短単位における品詞の割合（延べ語数）	POS	GVA_token_%_s	BCCWJ_token_%	PB_token_%_su	PM_token_%_su	PN_token_%_su	LB_token_%_su	OW_token_%_su	OT_token_%_su	OP_token_%_su	OB_token_%_su	OC_token_%_su	OY_token_%_su	OV_token_%_su	OL_token_%_su
	名詞	41.415	35.036	34.925	40.645	46.485	31.7	50.701	40.594	61.012	28.054	27.96	35.124	32.77	47.466
	助詞	27.184	30.043	30.523	28.541	27.012	31.762	24.011	29.706	18.159	33.024	32.294	28.954	30.541	24.382
	動詞	13.359	13.524	13.881	11.947	10.762	14.373	10.552	13.298	7.379	15.231	14.371	12.273	17.081	12.05
	助動詞	5.073	9.827	9.285	8.249	6.472	10.195	4.235	6.854	4.623	11.174	14.023	11.273	7.839	3.669
	接尾辞	5.393	3.199	3.219	3.204	4.518	2.991	5.349	3.474	4.911	2.527	2.344	2.951	2.311	4.132
	副詞	0.274	1.75	1.606	1.567	0.765	1.951	0.464	0.849	0.422	2.28	2.047	2.223	1.617	0.111
	形容詞	0.372	1.518	1.437	1.565	1.028	1.603	0.504	1.282	0.482	1.849	2.254	1.973	2.988	0.189
	代名詞	0.474	1.45	1.44	1.041	0.452	1.794	0.275	0.718	0.246	2.186	1.537	1.432	2.109	0.235
	形状詞	0.722	1.256	1.313	1.38	0.936	1.285	1.071	1.264	0.694	1.316	1.269	1.226	1.047	0.551
	連体詞	0.867	0.953	1.024	0.748	0.436	1.107	0.692	0.962	0.306	1.143	0.727	0.778	0.888	1.257
	接頭辞	2.288	0.83	0.718	0.744	0.93	0.635	1.222	0.578	1.407	0.626	0.804	0.969	0.445	4.088
	接続詞	2.574	0.46	0.496	0.286	0.179	0.432	0.924	0.384	0.326	0.384	0.257	0.383	0.209	1.87
	感動詞	0	0.155	0.132	0.084	0.024	0.173	0	0.037	0.034	0.208	0.114	0.442	0.155	0
	合計	100	100	100	100	100	100	100	100	100	100	100	100	100	100
短単位における品詞の割合（異なり語数）	POS	GVA_type_%_su	BCCWJ_type_%	PB_type_%_suw	PM_type_%_suw	PN_type_%_suw	LB_type_%_suw	OW_type_%_su	OT_type_%_suw	OP_type_%_suw	OB_type_%_suw	OC_type_%_suw	OY_type_%_suw	OV_type_%_suw	OL_type_%_suw
	名詞	82.761	90.858	89.432	86.337	86.268	89.234	88.575	82.833	86.74	82.873	85.594	87.03	72.224	84.469
	動詞	7.989	5.153	5.869	6.844	7.167	5.986	5.817	9.137	6.477	9.145	6.946	6.308	15.413	7.697
	副詞	1.234	1.659	1.785	2.285	1.645	1.881	0.879	2.009	1.781	2.766	2.83	2.564	3.296	0.666
	形状詞	1.883	0.912	1.146	1.697	1.676	1.14	1.398	1.692	1.482	1.911	1.638	1.477	2.221	1.684
	形容詞	0.48	0.429	0.532	0.802	0.808	0.547	0.573	0.972	0.687	1.035	0.824	0.735	1.922	0.607
	接尾辞	2.766	0.421	0.571	0.982	1.277	0.549	1.581	1.611	1.5	1.045	1.016	0.856	2.454	2.683
	感動詞	0	0.2	0.196	0.241	0.188	0.203	0.025	0.311	0.262	0.32	0.323	0.331	0.402	0
	接頭辞	1.299	0.145	0.196	0.31	0.397	0.186	0.569	0.53	0.489	0.327	0.338	0.283	0.668	0.979
	助詞	0.636	0.075	0.091	0.176	0.216	0.087	0.234	0.347	0.219	0.189	0.167	0.14	0.505	0.509
	代名詞	0.2338	0.065	0.077	0.123	0.108	0.082	0.083	0.197	0.104	0.157	0.127	0.107	0.32	0.118
	助動詞	0.285	0.038	0.047	0.092	0.118	0.048	0.104	0.179	0.12	0.103	0.088	0.075	0.271	0.294
	連体詞	0.22	0.026	0.036	0.066	0.079	0.035	0.09	0.099	0.078	0.079	0.068	0.057	0.163	0.137
	接続詞	0.21	0.019	0.024	0.046	0.053	0.024	0.072	0.084	0.059	0.05	0.041	0.037	0.141	0.157
	合計	100	100	100	100	100	100	100	100	100	100	100	100	100	100

使用corpus做EDA

		合同文章		出版図書	出版雑誌	出版新聞	図書館書籍	政府報告	課文	宣传文章	畅销書籍	Yahoo! 知识库	Yahoo! 博客	韵文	法律文章
		GVA収集	BCCWJ全体	各レジスター（頻度1までの見出し）											
		契約書データ（補助記号を入れて	PM～OWは統合 OT～OMは可変	出版書籍	出版雑誌	出版新聞	図書館書籍	特定目的白書	特定目的教科書	特定目的広報紙	特定目的ベスト	特定目的Yahoo!	特定目的Yhaoo!	特定目的韻文	特定目的法律
短単位における品詞の割合（延べ語数）	POS	GVA_token_%_	BCCWJ_token_	PB_token_%_su	PM_token_%_su	PN_token_%_su	LB_token_%_su	OW_token_%_s	OT_token_%_su	OP_token_%_su	OB_token_%_su	OC_token_%_su	OY_token_%_su	OV_token_%_su	OL_token_%_su
	名詞	41.415	35.036	34.925	40.645	46.485	31.7	50.701	40.594	61.012	28.054	27.96	35.124	32.77	47.466
	助詞	27.184	30.043	30.523	28.541	27.012	31.762	24.011	29.706	18.159	33.024	32.294	28.954	30.541	24.382
	動詞	13.359	13.524	13.881	11.947	10.762	14.373	10.552	13.298	7.379	15.231	14.371	12.273	17.081	12.05
	助動詞	5.073	9.827	9.285	8.249	6.472	10.195	4.235	6.854	4.623	11.174	14.023	11.273	7.839	3.669
	接尾辞	5.393	3.199	3.219	3.204	4.518	2.991	5.349	3.474	4.911	2.527	2.344	2.951	2.311	4.132
	副詞	0.274	1.75	1.606	1.567	0.765	1.951	0.464	0.849	0.422	2.28	2.047	2.223	1.617	0.111
	形容詞				1.565							2.254	1.973	2.988	0.189
	代名詞				1.041							1.537	1.432	2.109	0.235
	形状詞	0.722	1.256	1.313	1.38			1.071	1.264	0.694	1.316	1.269	1.226	1.047	0.551
	連体詞	0.867	0.953	1.024	0.748			0.962	0.962	0.306	1.143	0.727	0.778	0.888	1.257
	接頭辞	2.288	0.83	0.718	0.744			1.222	0.578	1.407	0.626	0.804	0.969	0.445	4.088
	接続詞	2.574	0.46	0.496	0.286	0.179							0.383	0.209	1.87
	感動詞	0	0.155	0.132	0.084								0.442	0.155	0
	合計	100	100	100	100	100	100	100	100	100	100	100	100	100	100
短単位における品詞の割合（異なり語数）	POS	GVA_type_%_su	BCCWJ_type_%	PB_type_%_suw	PM_type_%_suw	PN_type_%_suw	LB_type_%_suw	OW_type_%_su	OT_type_%_suw	OP_type_%_suw	OB_type_%_suw	OC_type_%_suw	OY_type_%_suw	OV_type_%_suw	OL_type_%_suw
	名詞	82.761	90.858	89.432	86.337	86.268							03	72.224	84.469
	動詞	7.989	5.153	5.869	6.844	7.167							308	15.413	7.697
	副詞	1.234	1.659	1.785	2.285	1.645	1.881	0.879	2.009	1.781	2.766	2.83	2.564	3.296	0.666
	形状詞	1.883	0.912	1.146	1.697	1.676	1.14	1.398	1.692	1.482	1.911	1.638	1.477	2.221	1.684
	形容詞	0.48	0.429	0.532	0.802	0.808	0.547	0.573	0.972	0.687	1.035	0.824	0.735	1.922	0.607
	接尾辞	2.766	0.421	0.571	0.982	1.277	0.549	1.581	1.611	1.5	1.045	1.016	0.856	2.454	2.683
	感動詞	0	0.2	0.196	0.241	0.188	0.203	0.025	0.311	0.262	0.32	0.323	0.331	0.402	0
	接頭辞	1.299	0.145	0.196	0.31	0.397	0.186	0.569	0.53	0.489	0.327	0.338	0.283	0.668	0.979
	助詞	0.636	0.075	0.091	0.176	0.216	0.087	0.234	0.347	0.219	0.189	0.167	0.14	0.505	0.509
	代名詞	0.2338	0.065	0.077	0.123	0.108	0.082	0.083	0.197	0.104	0.157	0.127	0.107	0.32	0.118
	助動詞	0.285	0.038	0.047	0.092	0.118	0.048	0.104	0.179	0.12	0.103	0.088	0.075	0.271	0.294
	連体詞	0.22	0.026	0.036	0.066	0.079	0.035	0.09	0.099	0.078	0.079	0.068	0.057	0.163	0.137
	接続詞	0.21	0.019	0.024	0.046	0.053	0.024	0.072	0.084	0.059	0.05	0.041	0.037	0.141	0.157
	合計	100	100	100	100	100	100	100	100	100	100	100	100	100	100

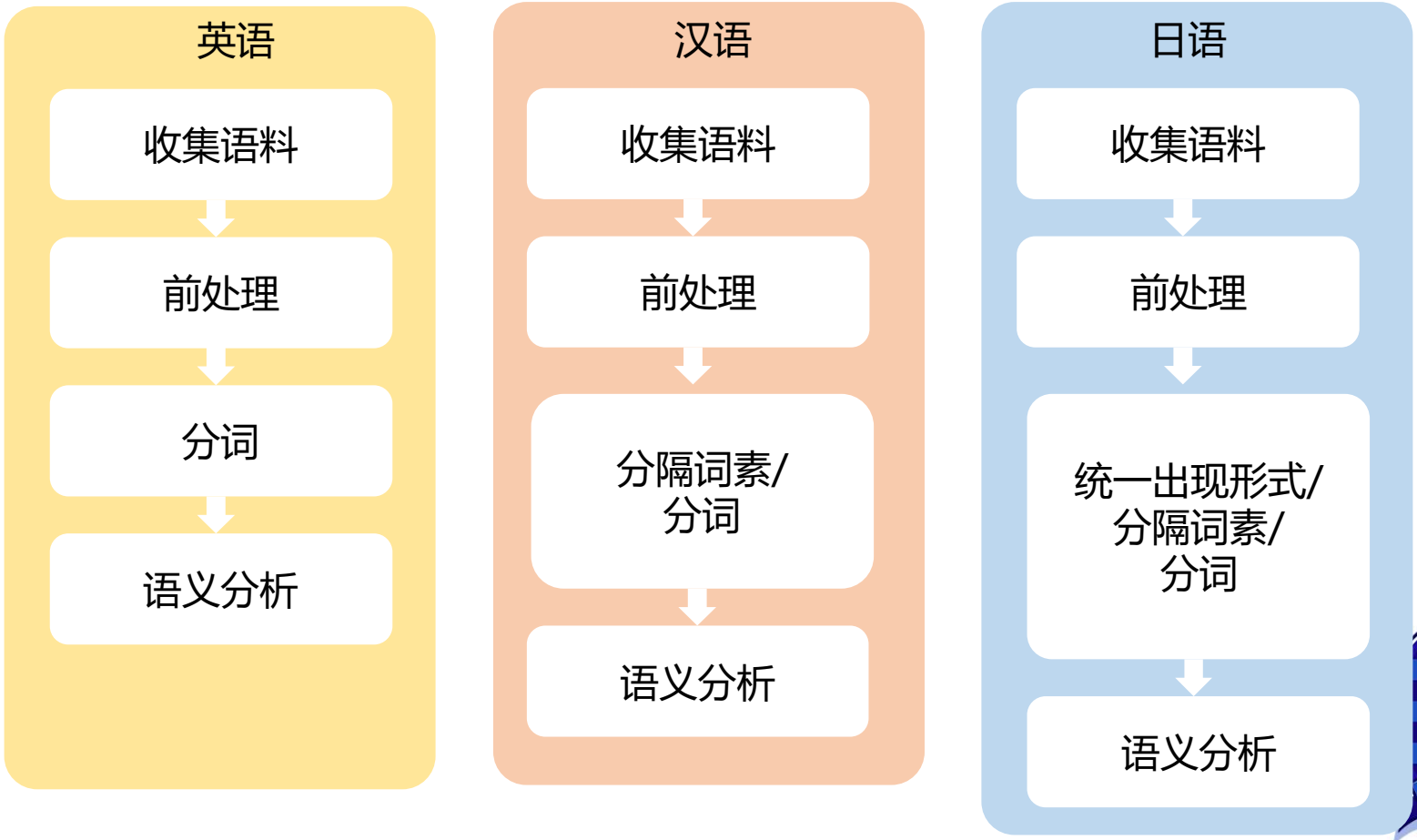
Q. 为什么要做EDA?

A. 为了把握自己的数据倾向。
这样可以:

- 后面的工作的处理的效率会变高
- 选model时有把握
- 对理解Model吐出来的结果有用

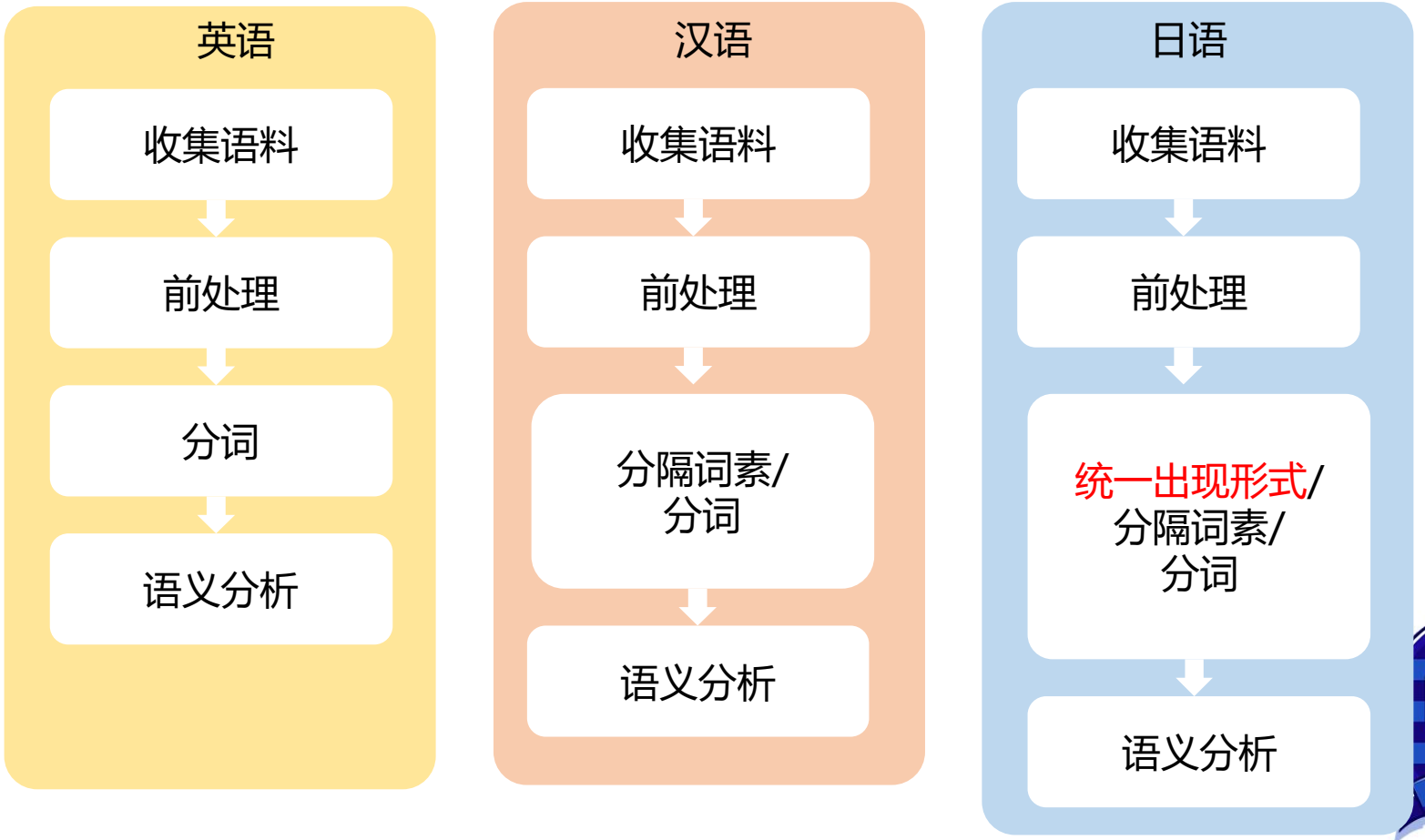
2 多语言NLP攻略

NLP基础@日中英


英语
收集语料
前处理
分词
语义分析
汉语
收集语料
前处理
分隔词素/
分词
语义分析
日语
收集语料
前处理
统一出现形式/
分隔词素/
分词
语义分析


PYTHON 30th

NLP基础@日中英


英语
收集语料
前处理
分词
语义分析
汉语
收集语料
前处理
分隔词素/
分词
语义分析
日语
收集语料
前处理
统一出现形式/
分隔词素/
分词
语义分析


PYTHON 30th

什么叫“统一出现形式”？

汉语

吃/了
没/吃

手机

日语

- 动词活用的统一工作

食べ/た
食べ/て/ない ➡ 食べる

- 名词的统一工作

携帯電話
携帯
ケータイ電話 ➡ 携帯電話
けいたい
ケータイ
ｹｰﾀｲ

NLP基础@日中英

英语

- 收集语料
- 前处理
- 分词
- 语义分析

汉语

- 收集语料
- 前处理
- 分隔词素/分词
- 语义分析

Mecab+neologd,
Mecab+unidic,
Juman
Ginza,
Janome

- 统一出现形式/分隔词素/分词
- 语义分析

NLP基础@日中英

英语

- 收集语料
- 前处理
- 分词
- 语义分析

汉语

- 收集语料
- 前处理
- 分隔词素/

日语

- 收集语料
- 前处理
- 统一出现形式/ 分隔词素/ 分词
- 语义分析

Ginza,
Cabocha,
COTOHA API

日中语义分析的不同点（之一）

汉语

太郎打花子。

施事 动作 受事

花子打太郎。

施事 动作 受事

日语

太郎は花子をぶった。

施事 受事 动作

花子は太郎をぶった。

施事 受事 动作

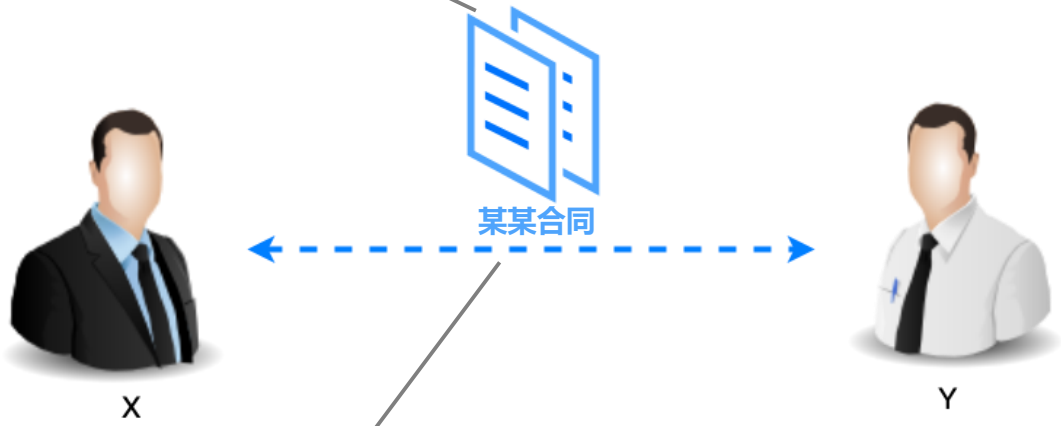
施事者和受事者都会有助词标记，
可以根据助词推测句法结构。

3 “合同风险预测模型”实战经验分享

什么叫“合同风险预测”？

合同内容有没有改写的条文没有写？

目的1：查看合同的条文类似性



合同内容有没有对自方/自方公司不利的条文？

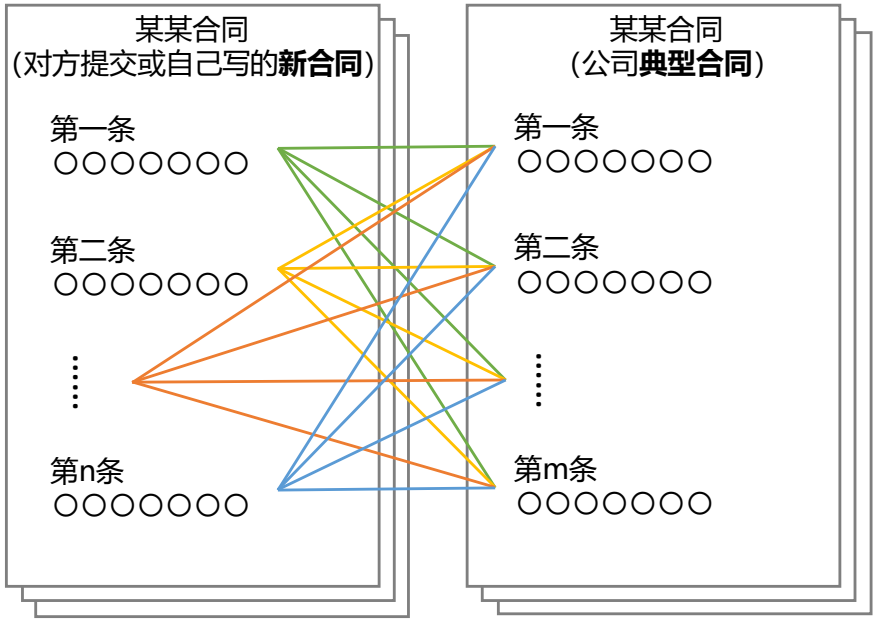
目的2：查看合同中的不利条文

※合同内容的对自方利不利根据各国的管辖法律

使用Python做“合同风险预测模型”的思路

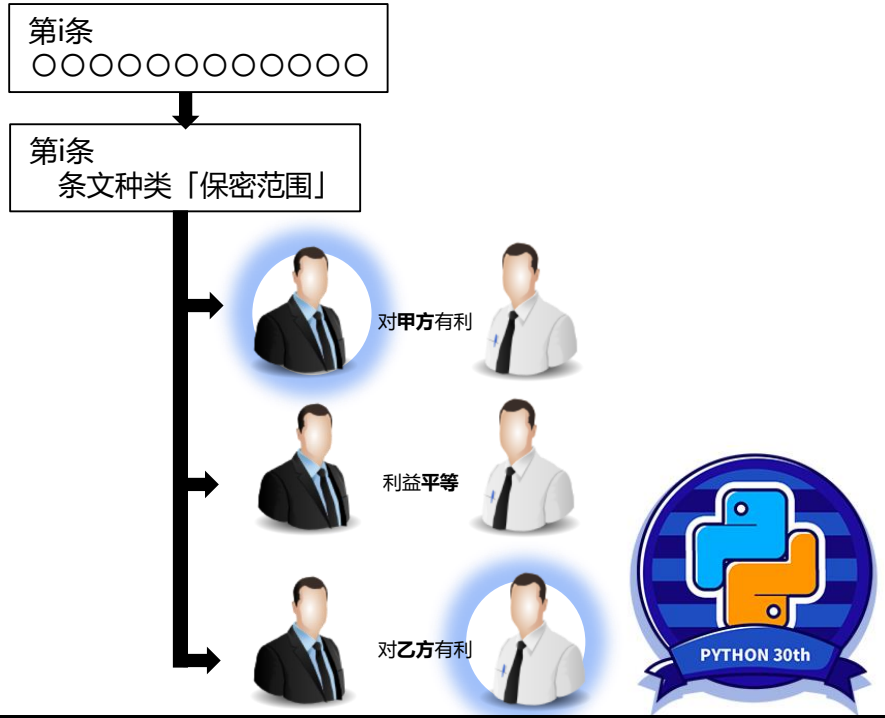
目的1：查看合同的条文类似性

- 需要算出条文和条文的类似度的值
- 根据Threshold（阈值）来判断条文内容的一致。



目的2：查看合同的不利条文

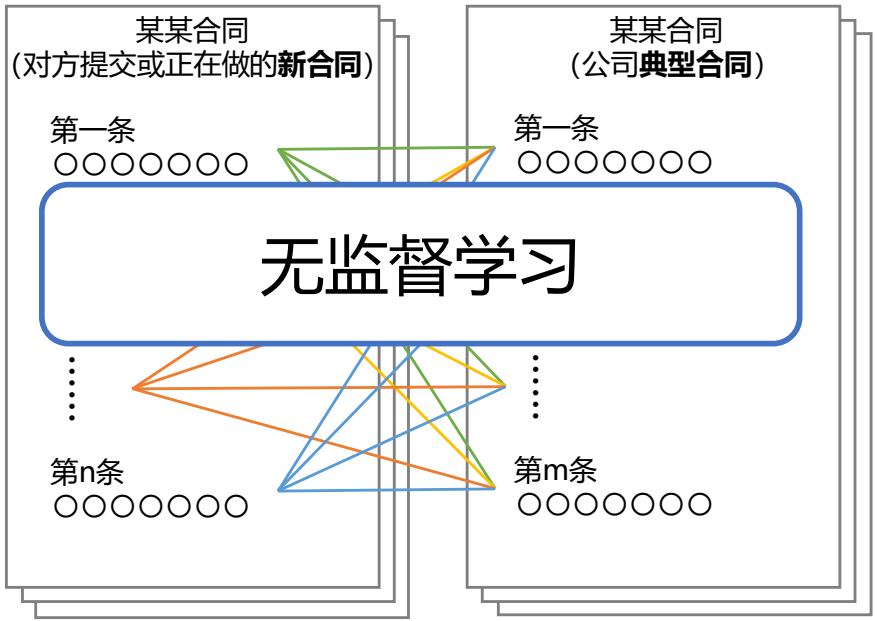
- 分类每一个条文的条文的种类（每个种类的风险都不一样）
- 根据条文的种类判断条文的内容的有利方。



使用Python做“合同风险预测模型”的思路

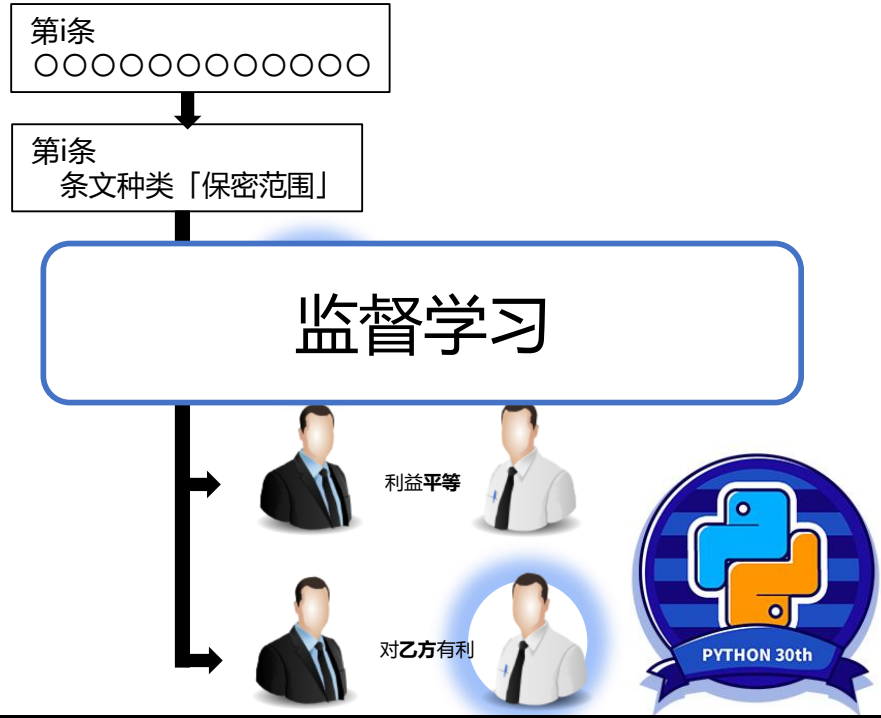
目的1：查看合同的条文类似性

- 需要算出条文和条文的类似度的值
- 根据Threshold（阈值）来判断条文内容的一致。



目的2：查看合同的不利条文

- 分类每一个条文的条文的种类（每个种类的风险都不一样）
- 根据条文的种类判断条文的内容的有利方。



预测合同的风险①

ROUGE

机器翻译的手法

```
$ from sumeval.metrics.rouge import RougeCalculator
$ rouge = RougeCalculator(lang='<填en/ja/zh>')
$ Rouge_L score = rouge.rouge_l(target, references)
```

word2vec

为什么不用doc2vec呢?

神经网络

```
$ from gensim.models import word2vec
$ model = word2vec.Word2Vec(input_data, size=<..>,
min_count=<..>, window=<..>)
```

RIBES

机器翻译的手法

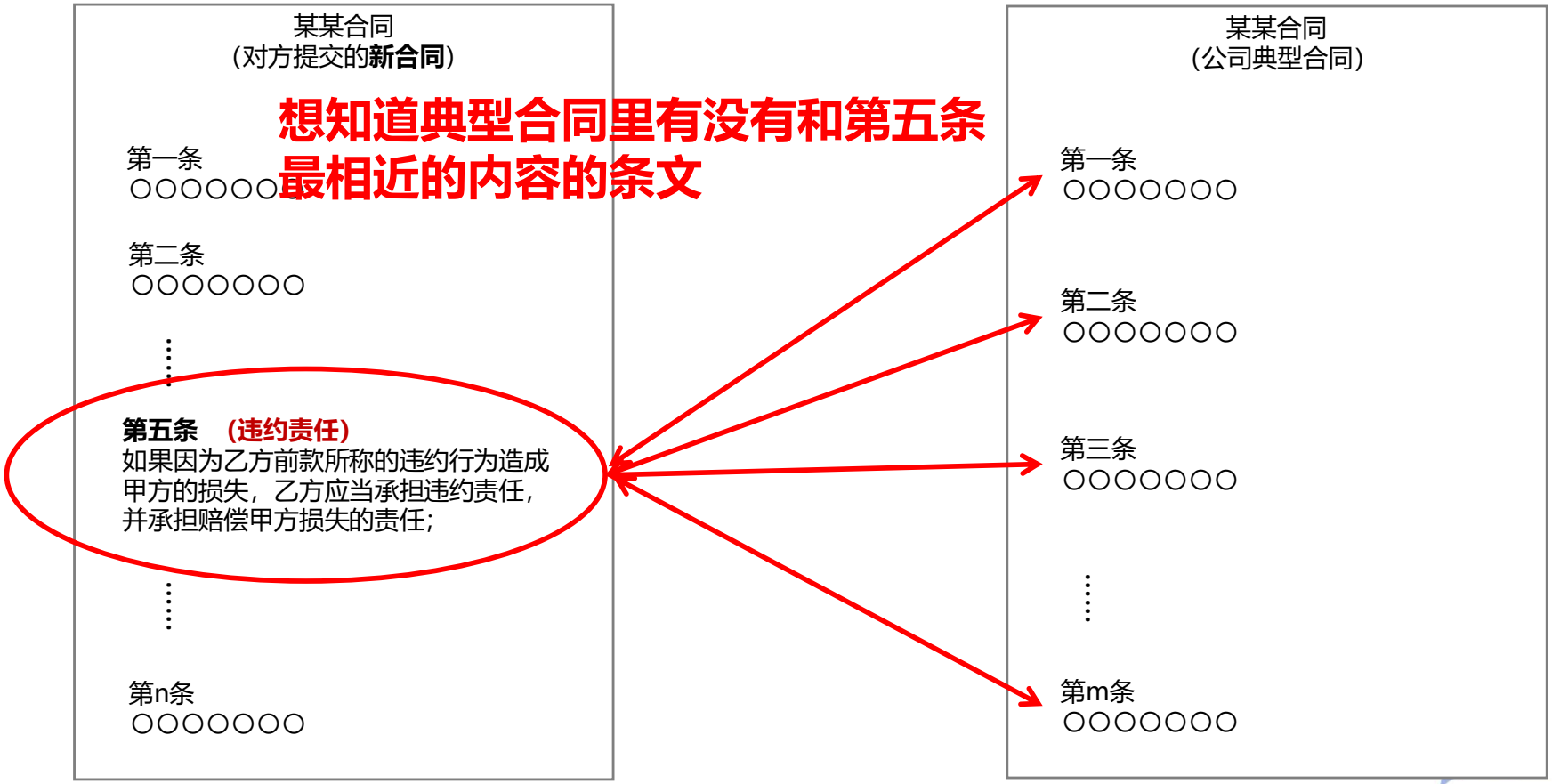
```
$ from RIBES import RIBESevaluator
$ ribes = RIBESevaluator()
$ score = ribes.eval([target], [[references]])
```

TF-IDF

向量化的典型手法

```
$ from sklearn.feature_extraction.text
import CountVectorizer
$ from sklearn.feature_extraction.text
import TfidfVectorizer
```

具体例子：


某某合同
(对方提交的新合同)
想知道典型合同里有没有和第五条最相近的内容的条文
第一条
〇〇〇〇〇〇〇
第二条
〇〇〇〇〇〇〇
第五条 (违约责任)
如果因为乙方前款所称的违约行为造成甲方的损失，乙方应当承担违约责任，并承担赔偿甲方损失的责任；
第n条
〇〇〇〇〇〇〇
某某合同
(公司典型合同)
第一条
〇〇〇〇〇〇〇
第二条
〇〇〇〇〇〇〇
第三条
〇〇〇〇〇〇〇
第m条
〇〇〇〇〇〇〇

条文的类似性结果比较

	条文	条文内容	ROUGE	RIBES	TF-IDF	Word2vec
第一条	甲乙双方确认：“秘密信息”是指甲方及其关联公司未曾公开的商业秘密、技术信息和财务信息等，包括但不限于设计、程序、制作工艺、制作方法、管理诀窍、产品或服务的销售网络、销售状况、客户名单、市场开发及售后服务情况、产销策略、招投标中的标底及标书内容。	秘密信息	0.0	0.0	0.0	0.088
	乙方承认在为甲方工作期间可能直接或间接地通过书面、口头、图表、音像资料等获得或通过观察全部或部分设备、产品等获得这些秘密信息。	秘密信息	0.133	0.0	0.309	0.304
	甲乙双方同意，上述“秘密信息”不包含那些非因乙方过错而进入公众领域的公开信息。	秘密信息	0.076	0.0	0.018	0.349
第二条	除履行职务需要之外，未经甲方事先书面同意，乙方不得泄漏、传播、公布、发表、传授、转让、交换或者以其他任何方式使任何第三方（包括无权知悉该项秘密的甲方职员）知悉属于甲方或者属于第三方但甲方承诺有保密义务的商业秘密，也不得在履行职务之外使用这些秘密信息。	保密责任	0.115	0.0	0.049	0.301
第三条	双方同意本协议规定的保密期限为自本协议签署之日起至双方劳动关系终止或解除后年内有效。	保密期限	0.0	0.0	0.0	0.006
	在保密期限内，乙方无论因何种原因从甲方或甲方关联公司离职，仍须承担如同任职期间一样的保密义务；乙方认可，甲方及甲方关联公司在支付工资报酬时，已考虑了乙方离职后需要承担的保密义务，故而无须在乙方离职时另外支付保密费。	保密期限	0.153	0.0	0.129	0.519
第四条	如果乙方违反本协议的规定，应赔偿甲方全部损失。赔偿范围包括但不限于甲方的名誉损失、直接损失和可得利益的损失，以及调查费用和诉讼费用、律师费用。	违约责任	0.235	0.58	0.268	0.757
	乙方违约后还应采取各种合理方法挽回泄密造成的影响，尽可能使秘密信息继续处于保密状态；同时，本协议继续有效。	违约责任	0.153	0.54	0.07	0.501

条文的类似性结果比较

	条文	条文内容	R	rd2vec
第一条	甲乙双方确认：“秘密信息”是指甲方及其关联公司未曾公开的商业秘密、技术信息和财务信息等，包括但不限于设计、程序、制作工艺、制作方法、管理诀窍、产品或服务的销售网络、销售状况、客户名单、市场开发及售后服务情况、产销策略、招投标中的标底及标书内容。	秘密信息		0.088
	乙方承认在为甲方工作期间可能直接或间接地通过书面、口头、图表、音像资料等获得或通过观察全部或部分设备、产品等获得这些秘密信息。	秘密信息		0.304
	甲乙双方同意，上述“秘密信息”不包含那些非因乙方过错而进入公众领域的公开信息。	秘密信息		0.349
第二条	除履行职务需要之外，未经甲方事先书面同意，乙方不得泄漏、传播、公布、发表、传授、转让、交换或者以其他任何方式使任何第三方（包括无权知悉该项秘密的甲方职员）知悉属于甲方或者属于第三方但甲方承诺有保密义务的商业秘密，也不得在履行职务之外使用这些秘密信息。	保密责任		0.301
第三条	双方同意本协议规定的保密期限为自本协议签署之日起至双方劳动关系终止或解除后 年内有效。	保密期限		0.006
	在保密期限内，乙方无论因何种原因从甲方或甲方关联公司离职，仍须承担如同任职期间一样的保密义务；乙方认可，甲方及甲方关联公司在支付工资报酬时，已考虑了	保密期限		0.519
第四条	如果乙方违反本协议的规定，应赔偿甲方全部损失。赔偿范围包括但不限于甲方的名誉损失、直接损失和可得利益的损失，以及调查费用和诉讼费用、律师费用。	违约责任		0.757
	乙方违约后还应采取各种合理方法挽回泄密造成的影响，尽可能使秘密信息继续处于保密状态；同时，本协议继续有效。	违约责任	0.153	0.501

最相近的条文

某某合同
（对方提交或正在做的**新合同**）

第一条
〇〇〇〇〇〇〇

第五条 **（违约责任）**
如果因为乙方前款所称的违约行为造成甲方的损失，乙方应当承担违约责任，并承担赔偿甲方损失的责任；

第n条
〇〇〇〇〇〇〇

word2vec需要注意的地方

- 最像“甲方”的是“乙方”0.922
- 普通情况下，可以说这次word2vec投进去的语料和parameter调的还不错。
- 但是，我们需要考虑合同文章的“甲”和“乙”利益相反的这一点。

<解决办法1>

选BERT等可以考虑到文章的前后关系的模型。（可是需要大量数据，还要考虑到可解释性的问题）

<解决办法2>

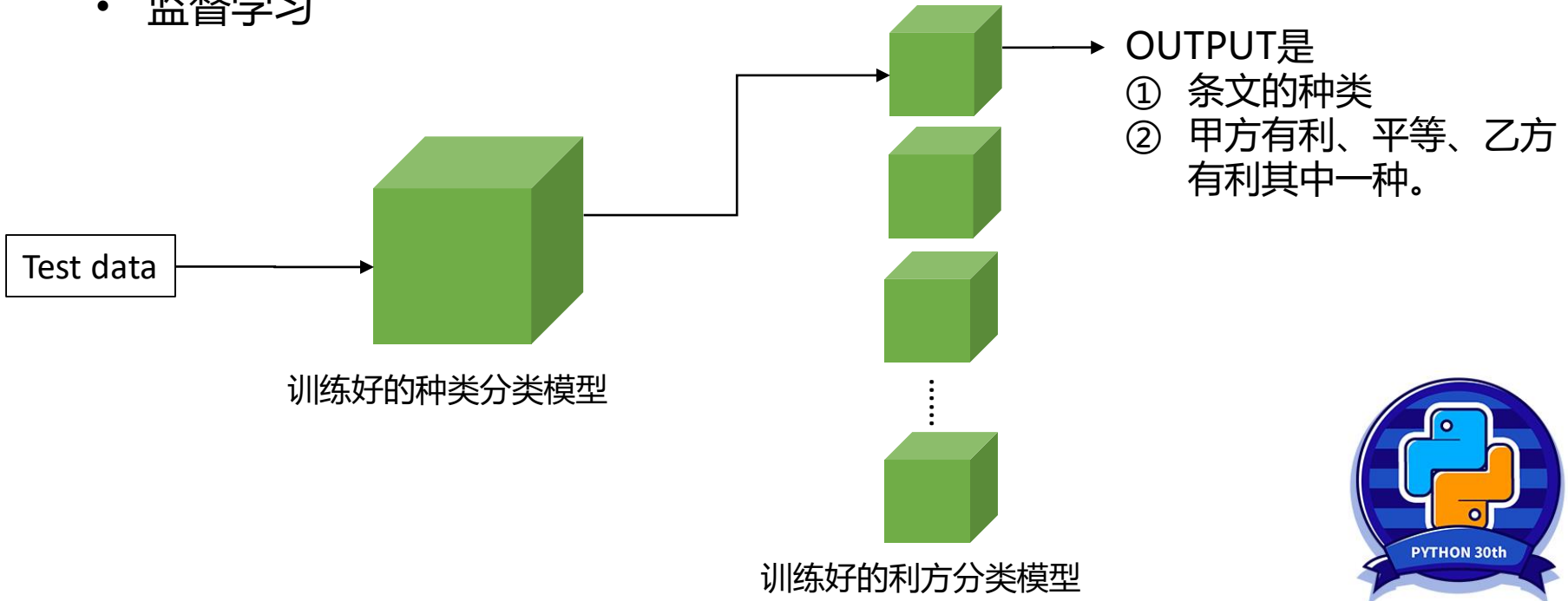
采取监督学习的模型，做分类问题

```
1  ret = model.wv.most_similar(positive=['甲方'])
2  for item in ret:
3      print(item[0], item[1])
```

```
乙方 0.9226242899894714
甲方乙方 0.5261456370353699
买方 0.4849575459957123
丙方 0.4754582643508911
其 0.46063581109046936
校方 0.4300388693809509
商户 0.41686224937438965
卖方 0.4086381196975708
无条件 0.40785565972328186
退货 0.4077020287513733
```

预测合同的风险②

- 使用RandomForest分类器(RF)做两种分类
  1. 预测条文的种类的multiclass分类
  2. 预测条文甲方有利/平等/乙方有利的multiclass分类
- 监督学习



4 总结

人生苦短，快去NLP。

内容总结

- 主要介绍了中文和日文的具体分析方法以及合同风险预测模型的思路和构建流程。

- 目前，日文的文的合同风险预测模型的结果是：

预测一致条文正确度达到**85%**
预测条文种类达到**91%**
预测利于何方正确度达到**90%**

总结

- 任何时候都不要忘了BUSIBESS上的课题是什么。
- 把课题掉入具体统计问题或能够使用机器学习解决的问题。
  （想不到应该怎么处理的时候，请先做EDA把握数据倾向）
- 解决问题的办法有很多，浅的、深的、单层的、深层的，根据你可以用的resource（时间和预算）来挑选分析方法以及模型构建。
- 目的确定了，模型选好了，剩下只需要你用Python去奔跑了！

THANK YOU

mina-zo

inazo18