算法 1：基于关键字的正样本资讯筛选算法

输入：关键字文件列表：pos_files[1]，neg_files[2]，pos_mixed_files[3]，neg_mixed_files[4]
待标记的资讯文件：$unlabel\_file$[5]

输出：正样例集合 $S_1 = \{news_1, news_2, \cdots, news_m\}$
负样例集合 $S_{-1} = \{news_1, news_2, \cdots, news_n\}$

算法步骤：

(1) 读取各个文件关键字:

$POS\_KEYS = read\_key\_files$[6] (pos_files)
$NEG\_KEYS = read\_key\_files(neg\_files)$
$POS\_MIXED\_KEYS = read\_key\_files(pos\_mixed\_files)$
$NEG\_MIXED\_KEYS = read\_key\_files(neg\_mixed\_files)$
$UNLABEL\_NEWS = read\_news\_file(unlabel\_file)$
$S_1 = \emptyset$, $S_{-1} = \emptyset$

(2) **FOR** 每一个 $NEWS \in UNLABEL\_NEWS$:

IF $contain\_key$[7]$(NEWS, NEG\_KEYS)$:
$S_{-1} = S_{-1} \cup NEWS$
ELSE IF $contain\_mixed\_key$[8]$(NEWS, POS\_MIXED\_KEYS)$:
$S_1 = S_1 \cup NEWS$
ELSE IF $contain\_mixed\_key(NEWS, NEG\_MIXED\_KEYS)$:
$S_{-1} = S_{-1} \cup NEWS$
ELSE IF $contain\_key(NEWS, POS\_KEYS)$:
$S_{-1} = S_{-1} \cup NEWS$
ELSE IF:
$S_{-1} = S_{-1} \cup NEWS$
END

---

[1] pos_files=['$keywords\_EVT.txt$', '$keywords\_GOV.txt$', '$keywords\_IDX.txt$', '$keywords\_VIP.txt$', '$keywords\_POS.txt$']
[2] neg_files = ['$keywords\_BLK.txt$', '$keywords\_ORG.txt$', '$keywords\_PER.txt$', '$keywords\_GPE.txt$', ''$keywords\_NEG.txt$'']
[3] pos_mixed_files = [' $mixed\_keywords\_POS.txt$ ']
[4] neg_mixed_files = ['$mixed\_keywords\_NEG.txt$ ']
[5] $unlabel\_file$: 需要标注资讯，文件格式：每行一条资讯
[6] 读取文件集函数，每个文件存储一类关键字。文件格式：一个关键字占一行，详见文件 keywords_EVT.txt
[7] 判断是否含有关键字。匹配规则：NEWS（单个）匹配到 KEYS 中一个关键字即为成功
[8] 逻辑同 7，主要区别在关键字组成：每个关键字由’+’连接，如’黄金’+‘美元’，必须同时匹配到才算成功。组合关键字组成详见文件$mixed_keywords_POS.txt$