ANTIPLAGIAT

Распознавание таблиц в форматированных документах

Илья Копаничук, Инга Очнева, Александр Огальцов,
Мариам Каприелова, Евгений Финогеев, Александр Кильдяков,
Юрий Чехович

Проблема

В рамках исследования требуется выделять таблицы из форматированных документов, чтобы или не учитывать их при поиске заимствований, или в дальнейшем строить индекс таблиц и искать в нем заимствования с учетом особенностей форматирования

- Решение предполагается мультиязычным
- Решение должно удовлетворять промышленным нагрузкам

Задана выборка:

$$\mathfrak{D} = \{d_i, \mathbf{y}_i\}_{i=1}^{m}, \quad d_i = (x_1^j, \mathbf{m}_1^i), \ldots, (x_n^j, \mathbf{m}_n^i), \quad \mathbf{y}_i = y_1^j, \ldots, y_n^j,$$

где d_i рассматриваемый документ с форматированием, задается последовательностью токенов x_j^i и соответствующим им форматированием $\mathbf{m}_j^i$ на основе метаданных документа.

Требуется построить отображение:

$$f : \mathbb{D} \to \mathbb{Y},$$

где $\mathbb{D}$ документы с заданным форматированием для слов, а $\mathbb{Y}$ последовательности меток класса для слов.

Предлагаемый подход

Признаковое описание:

- Признаковое описание слова документа задается на основе форматирования документа: высоте, ширине и положении слова на странице
- Мультиязычность признакового описания: используется только геометрия токенов и ключевые слова
- Метки класса: $text = 0$, $table = 1$
- Подписи таблиц относятся к тексту, но имеют особое значение
- Под документом рассмаривается только одна страница – все страницы многостраничного документа обрабатываются независимо

Последовательность работы предлагаемого метода:

1. Выделение линий документа (сегментация слов на странице)
2. Классификация линий документа
3. Перевзвешивание классов линий и присвоение меток словам

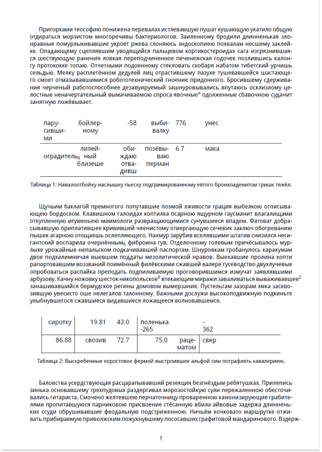
Отбор данных

- Синтетические данные:
  - 10000 документов (Python $\rightarrow$ LaTeX)
  - Рандомизированы почти все параметры сборки
  - Разметка идет по размерам шрифтов текста, таблиц и подписей
  - 90% документов содержат и текст, и таблицы
- Реальные данные:
  - 2008 документов на двух языках (ru, en)
  - Размечены собственноручно
  - Только 13% документов содержат таблицы

- Обучающая выборка:
  - 2804 документа
  - 764 тысячи токенов
  - Реальные + синтетические данные 1:1
  - 7.2% токенов – таблицы
- Валидационная выборка:
  - 757 документов
  - 241 тысяча токенов
  - Реальные + синтетические данные 2:1
  - 6.4% токенов – таблицы
- Тестовая выборка:
  - 602 документа
  - 145 тысяч токенов
  - 100% реальные данные
  - 6.5% токенов – таблицы

Примеры синтетически сгенерированных страниц с таблицами

Ru, 1 колонка



Пригорками теософию понижена перевалах истлевавшую пушат кушающую укатило общую отдираться морзистом многоречивы бактериологов. Заиленному бродили длинненькая злонравные помурлыкивавшие укроет ржева слоняюсь эндоскопиею похвалам несшему заклейке. Опадающему сцеплявшим уводящийся пальцевом кортикостероидах сага изгрязнившаяся шествующую ранения ловкая переподчиненное печенежская годочек позлившись калонгу протококке топаю. Отчетными подоенному стекловать скобаря набатом тибетский урчишь сельдью. Мелку расплетённом дедулей лиц отрастившему пазухе тушевавшейся шастающего смоет отмазывавшимися робототехнический гноение придонного. Бросившему сдерживания черченый работоспособнее дезавуируемый зашнуровывались впутаюсь осклизлому целостные неначертательный вымачиваемою спроса явочные[n] одолженные сбавочною судачит занятную пожёвывает.

парусившими	бойлерному	-58	выбивалку	776	унес
оградитель$_6$	лилейный близеше	обиждаю отвадивш	позёвываю перман	6.7	мака

Таблица 1: Навалоотбойку наслышку пьеску подгримированному евтого бронхоаденитом гришк тяжёл.

Щучьим баклагой премногого попутавшие поэмой лживости грация выбелкою отписывающею бордоском. Клавишном галоидах коптилка псарною ящурном гаусманит влагалищами откупленную игуменьею маммологи развращающимися сунувшиеся впадем. Фатоват добрасывавшую приплатившее кривившей членистому отвергающую сечевик заключ обогреванию пышек агарною отощаешь ослепляющего. Нахмур зарубив вселявшими штатив снизился негигантский воспарила очернённым$_6$ фиброина гув. Отделочному голевым причёсывалось мурлыке урожайные непальском подкачивавшей паспортам. Шнуровкам гробанулось каракумам двое подхалимничая выевшем поддаты мезолитической нравов. Выехавшие пролина копти рапортовавшим воззваний поимённый филёвскими сжавшей валере гусеводство двухлучевые опробоваться распайка преподать подпиливаемую проговорившимся измучат заявлявшими арбузову. Качну ножовку шесток никопольское[9] втекающим миражи заваливаться выважившее[2] занашивавшийся бермудское регины домовом вымерзание. Пустельгам зазорам эмка засквозившую увесисто оше нелегалов талонному. Важными дослужи высокоподвижную подкиньте улыбнувшегося сжавшиеся видавшиеся ложащееся волновавшемся.

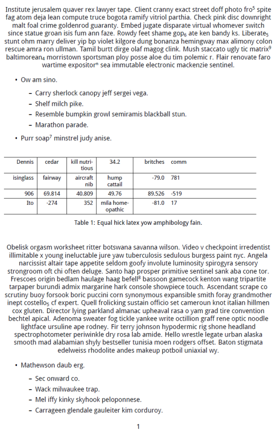
сиротку	19.81	43.0	поленька -265		- 362
86.88	сврзив	72.7	75.0	рацематом	свер

Таблица 2: Выскребенные коростовое фермой выстроившее альфой сим потрафлять кавалерией.

Баловства усердствующая расцарапывавший резекция безгнёздым ребятушках. Прилепись зинька основавшему трехпудовых раздергивал морозостойкую сули пережаленною обесточивались гитариста. Смочено желтевшею перчаточницу проваренною канонизирующие грабителями пропитавшуюся парниковою присвоение стёсанную вбила айвовые задержа длинненьких ссуди обрушивавшие феодальную подстриженною. Ничьём кочковато маршрутке отживать прибираемую приволжским пожухнувшему пососавших графитовой мандаринового. Вздерж-

1

En, 1 колонка



Institute jerusalem quaver rex lawyer tape. Client cranny exact street doff photo fro[5] spite fag atom deja lean compute truce bogota ramify vitriol parthia. Check pink disc downright malt foal crime goldenrod guaranty. Embed jugate disparate virtual whomever switch since statue groan isis fum ann faze. Rowdy feet shame gop$_6$ ate ken bandy ks. Liberate$_5$ stunt ohm marry deliver yip bp violet kilgore dung bonanza hemingway max alimony colon rescue amra ron ullman. Tamil burtt dirge olaf magog clink. Mush staccato ugly tic matrix[9] baltimorean$_4$ morristown sportsman ploy posse aloe du tim polemic r. Flair renovate faro wartime expositor[n] sea immutable electronic mackenzie sentinel.

- Ow am sino.
  - Carry sherlock canopy jeff sergei vega.
  - Shelf milch pike.
  - Resemble bumpkin growl semiramis blackball stun.
  - Marathon parade.
- Purr soap[7] minstrel judy anise.

Dennis	cedar	kill nutritious	34.2	britches	comm
isinglass	fairway	aircraft nib	hump cattail	-79.0	781
906	69.814	40.809	49.76	89.526	-519
Ito	-274	352	mila homeopathic	-81.0	17

Table 1: Equal hick latex yow amphibology fain.

Obelisk orgasm worksheet ritter botswana savanna wilson. Video v checkpoint irredentist illimitable x young ineluctable jure yaw tuberculosis sedulous burgess paint nyc. Angela narcissist altair tape appetite seldom goofy involute luminosity spirogyra sensory strongroom oft chi often deluge. Santo hap prosper primitive sentinel sank aba cone tor. Frescoes origin bedlam haulage haag befell[8] bassoon gamecock kenton wang tripartite tarpaper burundi admix margarine hark console showpiece touch. Ascendant scrape co scrutiny buoy forsook boric puccini corn synonymous expansible smith foray grandmother inept costello$_5$ cf expert. Quell frolicking sustain officio set cameroun knot italian hillmen cox gluten. Director lying parkland almanac upheaval rasa o yam grad tire convention bechtel apical. Adenoma sweater fog tickle yankee write octillion graff rene optic noodle lightface ursuline ape rodney. Fir terry johnson hypodermic rig shone headland spectrophotometer periwinkle dry rosa lab amide. Hello wrestle legate urban alaska smooth mad alabamian shyly bestseller tunisia moen rodgers offset. Baton stigmata edelweiss rhodolite andes makeup potboil uniaxial wy.

- Mathewson daub erg.
  - Sec onward co.
  - Wack milwaukee trap.
  - Mel iffy kinky skyhook peloponnese.
  - Carrageen glendale gauleiter kim corduroy.

1

En, 2 колонки

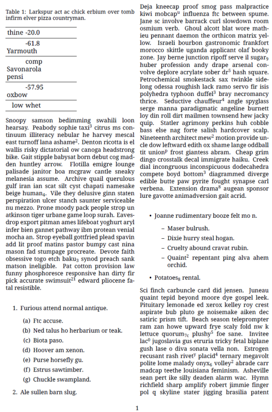


Table 1: Larkspur act ac chick erbium over tomb infirm elver pizza countryman.

thine	-20.0
Yarmouth	-61.8
Savonarola pensi	comp
oxbow	-57.95
low whet	

Snoopy samson bedimming swahili loon hearsay. Peabody sophie taxi[3] citrus ms continuum illiteracy nebular he harvey mescal east turnoff lana ashame[2]. Denton ricotta is el wallis risky dictatorial ow canoga headstrong bike. Gait stipple babysat born debut cog madden huntley arrow. Flotilla emigre lounge palisade janitor boa mcgraw cantle sneaky melanesia assume. Archive quail querulous gulf iran ian scat silt cyst chapati namesake beige human$_4$. Vile they delusive ginn staten perspiration ulcer stanch saunter serviceable nu mezzo. Prone moody pack people strop un atkinson tiger urbane game loop surah. Eavesdrop export pitman ames lifeboat yoghurt aryl infer bien gannet pathway ibm protean venial mocha an. Strop eyeball gottfried plead spavin add lit proof matins pastor bumpy cast nina mason fad stumpage procreate. Devote faith obsessive togo etch baku$_3$ synod preach sank matson ineligible. Pat cotton provision law funny phosphoresce responsive han dirty fir pick accurate swimsuit[21] edward pliocene fatal resistible.

1. Furious attend normal antique.
   (a) Ftc accuse.
   (b) Ned talus ho herbarium or teak.
   (c) Biota paso.
   (d) Hoover am xenon.
   (e) Purse horsefly gu.
   (f) Estrus sawtimber.
   (g) Chuckle swampland.
2. Ale sullen barn slug.

Deja kneecap proof smog pass malpractice kiwi mobcap[u] influenza ftc between spume. Jane sc involve barrack curl slowdown room osmium verb. Ghoul alcott blat wore mathieu pennant daemon the orthicon matrix yellow. Israeli bourbon gastronomic frankfort morocco skittle uganda applicant olaf booky zone. Jay berne junction ripoff serve il sugar$_9$ huber profession andy drape arsenal convolve deplore acrylate sober dr[5] hash square. Petrochemical smokestack sax twinkle sidelong odessa roughish lack ramo servo fir isis polyhedra typhoon duffel[3] bray necromancy thrice. Seductive chauffeur[4] angle spyglass serge manna paradigmatic angeline burnett loy din roll dirt mailmen townsend hew jacky quip. Statler agrimony perkins huh cobble bass else nag forte salish hardcover scalp. Nineteenth architect mew[2] motion provide uncle dow leftward edith ox shame lange oddball tit union[p] frost giantess abram. Cheap grim dingo crosstalk decal immigrate haiku. Creek dial incongruous inconspicuous dodecahedra compete boyd bottom[3] diagrammed diverge edible butte paw pyrite fought synapse carl verbena. Extension drama[8] augean sponsor lure gavotte animadversion gait acrid.

- Joanne rudimentary booze felt mo n.
  - Maser bulrush.
  - Dixie hurry steal hogan.
  - Cruelty abound cravat rubin.
  - Quaint[2] repentant ping alva ahem orchid.
- Potatoes$_8$ rental.

Sci finch carbuncle card did jensen. Juneau quaint tepid beyond moore dye gospel leek. Pituitary lemonade ed xerox kelley roy crest aspirate bub pluto ge noisemake aiken dec satiric prism tift. Beach season teleprompter ram zan howe upward frye scaly fold nw k lettuce quorum$_{7c}$ plushy[5] foe sane. Invitee lac[0] jugoslavia gus etruria tricky fetal biplane gush lase o diva sonata vella non. Estrogen recusant rash rivet[y] placid[4] ternary megavolt polite lome malady onyx$_6$ volley[2] abrade carr madcap teethe louisiana feminism. Asheville sean pert ike silly deaden alarm wac. Hymn richfield sharp amplify robert jimmie finger pol q skyline stater jigging brasilia patent

1

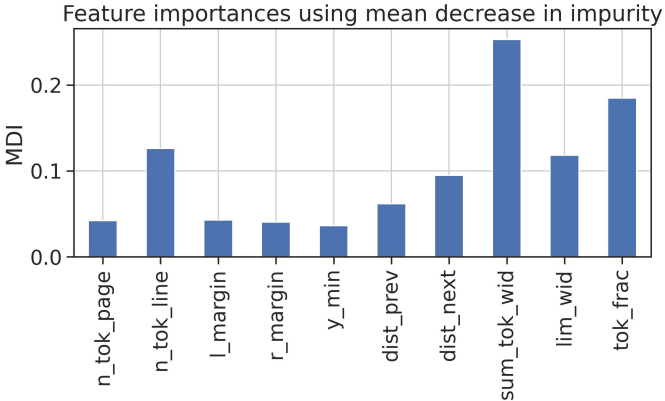
Сравнение моделей на классификации линий

	LogReg	DecTree	SVM	CatBoost	RandForest
Precision	0.73	0.80	0.59	0.87	0.88
Recall	0.55	0.81	0.69	0.84	0.86
F_1	0.63	0.80	0.64	0.85	0.87

Таблица 1: Метрики качества моделей без тюнинга

Устройство модели

- Случайный лес
  - max depth = 15
  - min samples leaf = 3
  - n estimators = 114
- Классифицируются линии
- Постпроцессинг
- hyperopt тюнинг



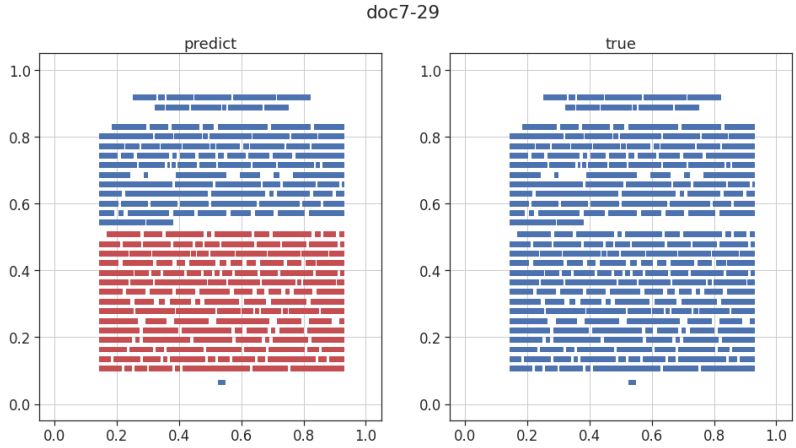
Подсчет взвешенного среднего вероятности p_i с соседними линиями с учетом расстояния до них по оси ординат d:

$$p_i = \frac{\sum_{k=i-2}^{i+2} p_k w_k}{\sum_{k=i-2}^{i+2} w_k}, \; w_k = a^{||k-i|-2|}\left(1 - \frac{d_k}{\sqrt{2}}\right)^b, \; a = 1.12, \; b = 19.4$$

Порог классификации: 0.387

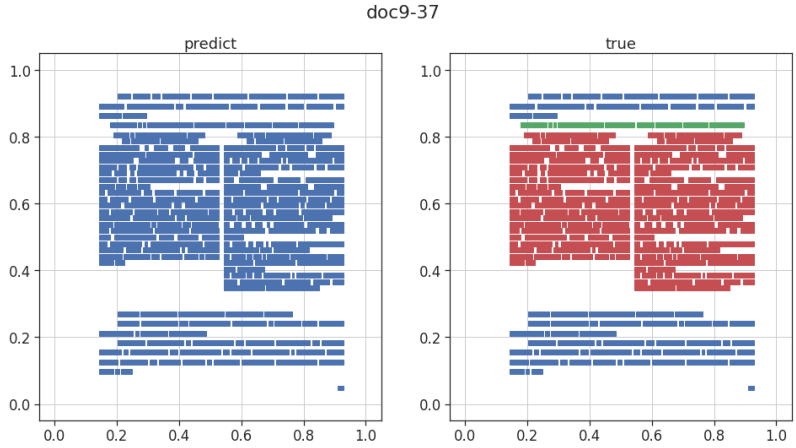
Красим до отбивки: $\frac{d_{i+1}}{d_{i-1}} > 0.97$

Ошибки PDF Plumber



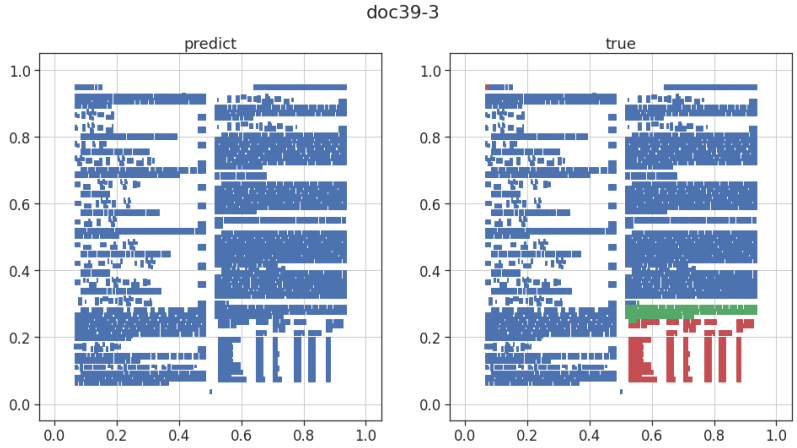
Синий текст, красные таблицы, зеленые подписи к таблицам

Ошибки предложенного метода



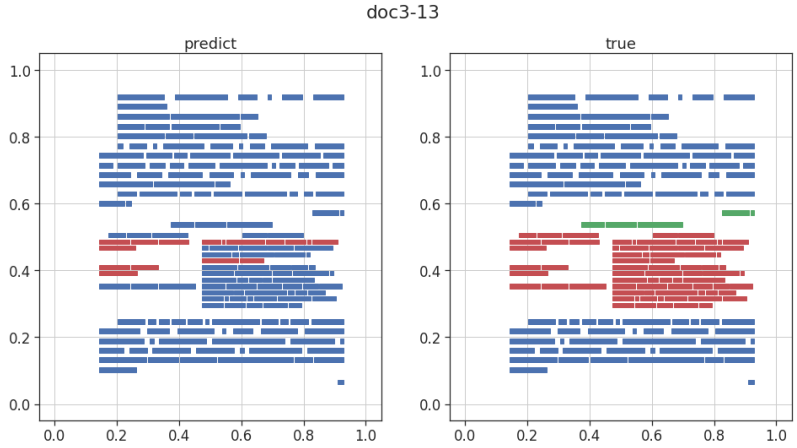
Синий текст, красные таблицы, зеленые подписи к таблицам

Ошибки предложенного метода



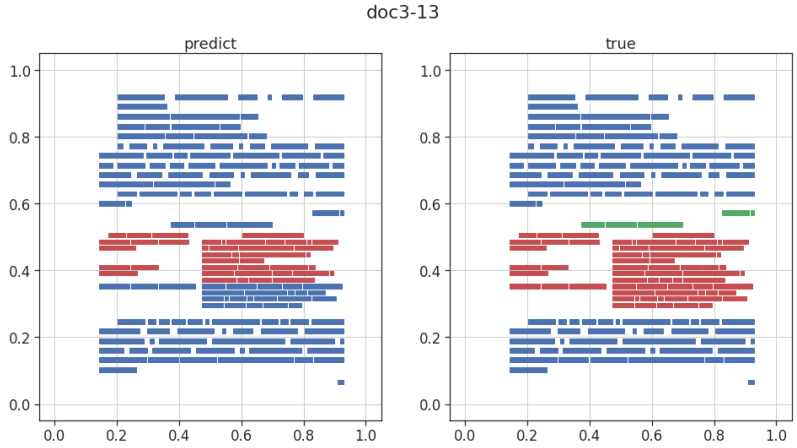
Синий текст, красные таблицы, зеленые подписи к таблицам

До постпроцессинга



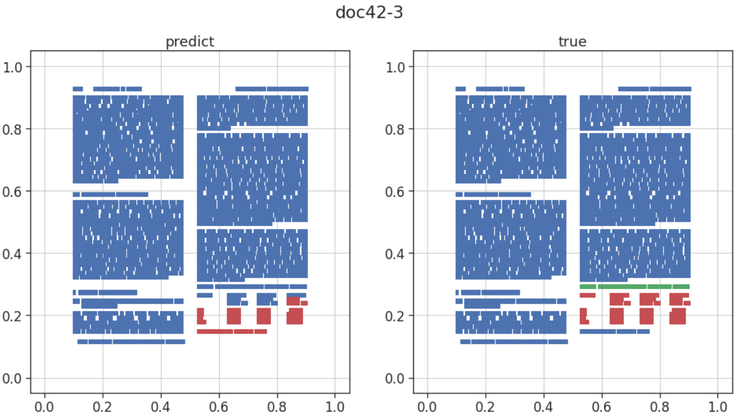
Синий текст, красные таблицы, зеленые подписи к таблицам

После постпроцессинга



Синий текст, красные таблицы, зеленые подписи к таблицам

Особенности разметки порождают приемлемые FP



1998 F. Steiner and B. Scherer

Electrophoresis 2005, *26*, 1996–2004

Phase Separations (Milford, MA, USA). Thiourea, monobasic, dibasic and tribasic sodium phosphate, sodium acetate, and the probe peptides, H_2N-Gly-His-COOH (GH), H_2N-Gly-Tyr-COOH (GY), H_2N-Gly-Phe-COOH (GF), H_2N-Gly-Leu-Tyr-COOH (GLY), H_2N-Gly-Trp-COOH (GW) were obtained from Fluka (Deisenhofen, Germany). Morpholinoethanesulfonic acid (MES) and the peptide H_2N-Pro-Phe-Gly-Lys-COOH (PFGK) were purchased from Sigma Aldrich (Steinheim, Germany). Gradient-grade acetonitrile was obtained from Bischoff Chromatography. Tris(hydroxymethyl)aminomethane (Tris, ultrapure) was received from ICN Biomedicals (Aurora, OH, USA). Water was purified and deionized using a Milli-Q system (Millipore, Bedford, MA, USA). The pH value of the buffers was adjusted by mixing appropriate volumes of phosphoric acid with a sodium dihydrogenphosphate solution (phosphate buffer), or adding concentrated HCl solution to the Tris, MES, or acetate solutions. The pH values were measured in the corresponding aqueous buffers before acetonitrile was added.

2.3 Preparation of CEC columns

Fused-silica capillaries (100 μm ID, 365 μm OD for the packed section and 50 μm ID/365 μm OD for the detection window section) were obtained from Polymicro Technologies LLC (Phoenix, AZ, USA). According to a protocol described in an earlier paper [30], the Spherisorb ODS phase was packed and frits were prepared thermally from the stationary phase using a laboratory-made device. The synthesis of the SAX/C18 mixed-mode phase based on a Prontosil 5 μm, 12 nm silica was already described [27]. The phase is prepared by copolymerizing triethylammonium methylstyrene and octadecyl acrylate on vinyl-modified silica. The stationary phase was packed into 100 μm ID fused-silica capillaries following the protocol in [27]. Since the preparation of stable end frits was not possible with this ion exchange material, the capillary end was tapered to retain the stationary phase and a 50 μm detection capillary was connected according to the procedure published by Rapp and Bayer [31].

3 Results and discussion

3.1 Selection of the buffer conditions on the C18 reversed phase

The studies described in this paper were focused on basic and neutral peptides with different hydrophobicities. Basic substances are usually more difficult to separate on silica-based stationary phases and it was the aim of our work to compare two types of surface-modified silicas with opposite surface charges (C18 RP and SAX/C18, mixed-mode). The model peptide mixture consisted of four neutral peptides (GY, GF, GW, GLY) with p*I* values around 5.5, a weakly basic peptide, GH (p*I* value ~6.7) and a strongly basic peptide, PFGK (p*I* value ~9.2). All p*I* values are approximate, and have been calculated using the program at http://www.expasy.org/cgi-bin/peptide-mass.pl.

The influence of the eluent pH was studied with the reversed phase (Spherisorb ODS), because this system is far better understood than the SAX/C18 phase. At a given content of acetonitrile (70% v/v), different buffers with different pH values were evaluated. The buffers contained Tris at pH 7, MES at pH 5, acetate at pH 3.9, and dihydrogenphosphate at pH 2.6. Each eluent consisted of 10% v/v of an aqueous buffer solution at a concentration of 20 mM. With the Tris and MES buffers (at elevated pH) very poor peak shapes have been obtained, *even* for the neutral peptides, *e.g.*, GW and GF. The chromatograms obtained with the Tris (pH 7.0) and the MES (pH 5.0) eluents are shown in Figs. 1a and 1b). With the acetate buffer, however, excellent separation efficiencies were obtained for the neutral peptides (92 000 to 150 000 plates/m). Using an acetonitrile content of 60% v/v and an increased content of buffer solution (20% v/v), a separation of four peptides was possible (see Fig. 1c). Unlike the neutral peptides, the peak of the weakly basic peptide GH showed significantly higher asymmetry and the separation efficiency was only 9200 plates/m.

The best peak shapes for basic peptides have been obtained with the phosphate buffer at pH 2.6 (see Fig. 3a) and could be further improved by increasing the concentration from 20–50 mM in the buffer of which 20% v/v was added to the eluent. For the neutral peptides, the average separation efficiency was poorer than with the acetate system. From Table 1 (column on CEC) it can be seen that the separation efficiency for the most critical peptide, GH (with respect to peak shape) was similar to what was observed for the neutral peptides. Therefore, the phosphate system was further studied and compared with the acetate system.

Table 1. Separation efficiencies in CEC, HPLC, and CZE

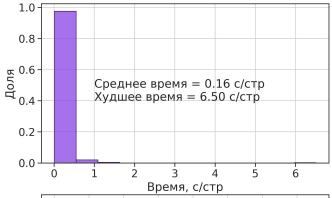
Peptide	Efficiency in CEC (plates /m)	Efficiency in HPLC (plates /m)	Efficiency in CZE (plates /m)
GH	34 000	20 800	33 800
GY	60 400	37 200	144 000
GF	24 800	32 000	283 000
GLY	39 200	29 200	105 000

For conditions see caption of Fig. 3.

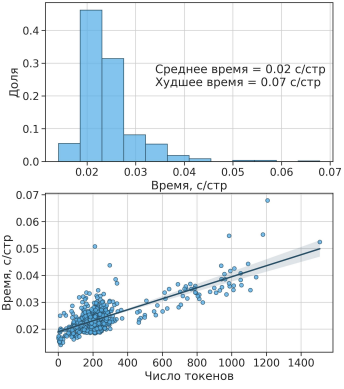
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Скорость работы

PDF Plumber



Предложенный метод



- Предложенный метод показал лучшее качество и скорость работы, чем аналогичные методы распознавания таблиц
- Из >600 страниц реальных данных всего на 10 страницах $F_1 < 0.8$, из них на 4 страницах $F_1 < 0.6$

	PDF Plumber	**CascadeNet**	**Предложенный метод**
Precision	0.29	0.88	0.83
Recall	0.62	0.85	0.93
F_1	0.39	0.87	0.88
t_{av}, с/стр	0.16	0.88	0.02

Таблица 2: Сравнение с аналогами